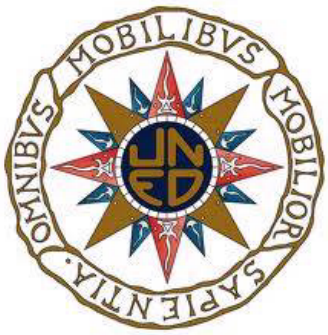


Universidad Nacional de Educación a Distancia

Intrinsic Semantic Spaces for the representation of documents and semantic annotated data

Thesis submitted by

Juan José Lastra Díaz

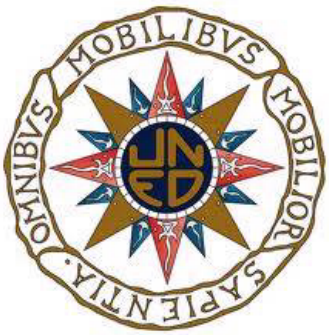
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE

Under the supervision of

Dra. Ana García Serrano

Defended in Madrid on September 29, 2014
Publicly available on December 24, 2014

This page intentionally left blank.



Universidad Nacional de Educación a Distancia

Máster Universitario en Lenguajes y Sistemas Informáticos
Especialidad: "Tecnologías del Lenguaje en la Web"
Departamento de Lenguajes y Sistemas Informáticos
Escuela Técnica Superior de Ingeniería Informática

Intrinsic Semantic Spaces for the representation of documents and semantic annotated data

MSc Thesis

MSc student:
Juan José Lastra Díaz

Thesis advisor:
Dra. Ana García Serrano

Defended in Madrid on September 29, 2014
Publicly available on December 24, 2014

This page intentionally left blank.

A Fátima, mi maravillosa,
amada y paciente esposa,
por haber soportado
estoicamente mi dedicación
a la etapa que culmina
con este trabajo.

A mis queridos hijos,
Fátima, Iñigo y Jaime,
a quienes he escatimado
algo de dedicación
durante estos años.

Contents

List of Tables

List of Figures

Abstract

This thesis introduces two novel semantic representation spaces for text documents and semantically annotated data, which are based in an intrinsic geometry approach, as well as other results, among which we have: (1) a novel ontology-based semantic distance, that we call *weighted Jiang-Conrath*, and (2) generalized normal distribution on differential manifolds, called *geodesic normal distribution*, what lead us to the definition of the *geodesic Mahalanobis distance.* By last, we prove that any Bayes classifier on a manifold defines a dual Voronoi diagram on it.

The ontology-based IR model looks promising, but it has not been evaluated experimentally yet. By other hand, the text document classifier yielded a first discouraging result due to the difficulties for the training of the model.

The common thread of our research is the use of notions of intrinsic differential geometry and geometric invariance, as means to bridge some gaps in the literature. The ontology-based IR model, as well as the text classifier proposed in this thesis, is inspired by a geometric approach, whose core idea is the integration of the geometric structures of the problem in the semantic representation spaces of the information. In summary, our approach attempts to build better models of semantic spaces by incorporating the properties and constraints of the mathematical objects involved in its definition.

The first part of the thesis introduces a novel ontology-based IR model based in a structure-preserving embedding of a populated ontology into a metric space that we call *Intrinsic Ontological Spaces*. The second part of the thesis introduces a novel text classifier, called *Intrinsic Bayes-Voronoi*, which is based in the representation of the document vectors by a manifold-based generative model, where the distribution function is defined on the unit hypersphere, instead of the euclidean ambient space.

The *Intrinsic Ontological Spaces* introduces a novel theoretical IR model that looks promising, although it has not even been evaluated experimentally. The proposed IR model is described in depth and validated with regard to our design axioms.

The motivation behind of our model is the finding of a set of geometric inconsistencies in some ontology-based IR models in the literature, which are derived from certain overlooked properties in their adaptations of the Vector Space Model (VSM). In essence, *our model refutes the unreflective use of the VSM model* in the fields of natural language processing (NLP) and information retrieval (IR).

Despite that the theoretical approach is interesting by itself, our main hypothesis is that the structure-preserving approach proposed by our model, should lead us to improve the quality of the ranking, as well as the measures of precision and recall in the semantic information retrieval systems.

The *Intrinsic Ontological Spaces* are, up to our knowledge, the first ontology-

based IR model to build a whole ontology-based structure-preserving representation for any sort of semantically annotated data in a populated ontology. In our model, every component has been designed with the aim to preserve the intrinsic geometry of any base ontology. The intrinsic geometry of any ontology is defined by three algebraic structures: (1) the order relation of the taxonomy, (2) the set inclusion relation, and (3) its intrinsic semantic metric. In this way, the methods for the representation of the queries, information units, weighting, ranking and retrieval, have been designed from geometric principled-based axioms, with the aim to capture all the semantic knowledge encoded in the base ontology. Using the language of the theory of categories, our model builds a natural equivalence, or morphism, among the input populated ontology and the representation space for the indexed information units.

Finally, the classifier of Bayes-Voronoi, introduced in the second part of the thesis, uses a manifold-based generative model to represent documents which is defined by a vector normal distribution on the unit hypershere, and we have called *geodesic normal distribution*. The distribution is defined on the unit hypershere, considered as a manifold, instead of the ambient space. The core idea is the observation that the normalized vectors are defined on the unit hypersphere, instead of the whole euclidean ambient space, and the proposed model explicitly integrates this constraint. The model removes one dimension to the normalized vectors, which corresponds to the projection of the data vectors on the unit hypershere (normalization). The geodesic normal distribution lead us to the definition of the Mahalanobis distance on a differential manifold, distance that we call *geodesic Mahalanobis distance.* We also prove that any Bayes classifier on a manifold defines a dual Voronoi diagram on it.

Keywords: ontology-based IR models, ontology-based semantic distances, semantic information retrieval, taxonomic semantic spaces, vector semantic spaces, semantic distances, Jiang-Conrath distance, valuation metrics, geodesic Mahalanobis distance, Hausdorff distance, semantic metric spaces, manifold-based distribution, text classifier.

Resumen

Esta tésis presenta dos nuevos espacios de representación semántica para documentos de texto y datos anotados semánticamente, los cuales se basan en un enfoque de geometría intrínseca, así como otros resultados, entre los cuales tenemos: una nueva distancia semántica sobre ontologías denominada *distancia ponderada de Jiang-Conrath*, una distribución normal generalizada sobre variades diferenciables que denominamos *distribución normal geodésica*, la cual nos conduce a la definición de la *distancia geodésica de Mahalanobis*. Por último, probamos que cualquier clasificador de Bayes sobre una variedad induce un diagrama de Voronoi dual sobre su dominio.

El modelo de recuperación de la información (RI) basado en ontologías parece prometedor, a pesar de aún no haber sido evaluado experimentalmente. Por otro lado, el clasificador de texto ha arrojado un primer resultado desesperanzador debido a ciertas dificultades en el entrenamiento del modelo.

El hilo conductor de nuestra investigación es el uso de nociones de geometría diferencial e invarianza geométrica como medio para cubrir algunas oprtunidades de mejora y problemas encontrados en los modelos actuales encontrados en la bibliografía. Tanto el modelo RI basado en ontologías, como el clasificador de texto propuestos en esta tésis, son inspirados por un enfoque geométrico, cuya principal idea es la integración de las estructuras geométricas del problema en los espacios de representación semántica de la información. En suma, nuestro enfoque intenta construir mejores modelos de espacios semánticos mediante la incorporación de las propiedades y restricciones de los objetos matemáticos involucrados en su definición.

La primera parte de la tésis presenta un nuevo modelo RI basado en ontologías que define una inmersión de una ontología poblada en un espacio métrico, la cual es denominada *Espacios Intrínsecos Ontológicos* y tiene como principal propiedad la preservación de las estructuras codificadas en las ontologías. En la segunda parte, presentamos un nuevo clasificador de documentos de texto, denominado *Clasificador Intrínseco de Bayes-Voronoi*, el cual se basa en la representación de los vectores de documento mediante un modelo generativo expresado sobre una variedad diferenciable, cuya función de distribución es definida sobre la hiperesfera unitaria, en vez de sobre el espacio euclídeo ambiente.

Los *Espacios Intrínsecos Ontológicos* introducen un nuevo modelo téorico de recuperación de la información que parece prometedor, si bien, como ya hemos señalado, éste aún no ha sido evaluado experimentalmente. El modelo propuesto es descrito en profundidad y validado con respecto a nuestros axiomas de diseño.

La motivación detrás de nuestro modelo es el descubrimiento de un conjunto de inconsistencias geométricas en algunos modelos RI basados en ontologías, las cuales

se derivan de ciertas propiedades pasadas por alto en sus adaptaciones del modelo de espacio vectorial (VSM). En esencia, *nuestro modelo refuta el uso irreflexivo del modelo VSM* en toda clase de tareas semánticas en el ámbito del procesamiento natural del lenguaje.

A pesar de que el enfoque teórico es intersante por sí mismo, nuestra principal hipótesis es que el enfoque invariante que proponemos en el modelo, debería conducirnos a mejorar la calidad de clasificación, así como las medidas de precisión y cobertura en los sistemas de recuperación de información de tipo semántico.

Los *Espacios Intrínsecos Ontológicos* son, hasta donde alcanza nuestro conocimiento, el primer modelo de recuperación de la información basado en ontologías donde cada componente del sistema ha sido diseñado basado en la ontología base para preservar todas las estructuras intrínsecas presentes. La geometría intrínseca de una ontología es definida por tres estructuras algebraicas: (1) la relación de orden de la taxonomía, (2) la relación de inclusión de conjuntos, y (3) su métrica semántica intrínseca. De esta forma, los métodos para la representación de las consultas, unidades de información, funciones de pesado, clasificación por relevancia y recuperación, han sido diseñados a partir de axiomas fundamentados en principios geométricos, con el objetivo de capturar todo el conocimiento codificado en la ontología base. Empleando el lenguaje de la teoría de categorías, nuestro modelo construye una equivalencia natural, o morfismo, entre la ontología poblada de entrada y el espacio de representación para las unidades de información indexadas.

Finally, ell clasificador de Bayes-Voronoi, introducido en la segunda parte de la tésis, emplea un modelo generativo para representar documentos de texto, el cual es definido por una distribución normal vectorial sobre la hiperesfera unitaria, la cual denominamos *distribución normal geodésica*. Dicha distribución es definida sobre la hiperesfera unitaria, vista como una variedad diferenciable, en vez de sobre el espacio euclídeo ambiente. La idea clave es la observación de que los vectores normalizados están contenidos en la hiperesfera unitaria, en de vez de sobre el espacio euclídeo ambiente, y el modelo propuesto integra de forma explícita dicha propiedad. El modelo reduce una unidad la dimensión de los vectores normalizados, la cual corresponde a la proyección de los vectores de datos sobre la hiperesfera unitaria (normalización). Asimismo, la distribución normal geodésica nos conduce a la definición de la distancia de Mahalanobis sobre una variedad diferenciable, distancia que denominamos *distancia geodésica de Mahalanobis.* Por último, probamos que cualquier clasificador de Bayes sobre una variedad induce un diagrama de Voronoi dual sobre su dominio.

Keywords: ontology-based IR models, ontology-based semantic distances, semantic information retrieval, taxonomic semantic spaces, vector semantic spaces, semantic distances, Jiang-Conrath distance, valuation metrics, geodesic Mahalanobis distance, Hausdorff distance, semantic metric spaces, manifold-based distribution, text classifier.

Acknowledgements

This thesis is the completion of three academic years in part-time mode with a huge personal effort and devotion, as most of UNED students, which would have not been possible without the support of my devoted wife Fátima, and my three children: Fátima, Iñigo and Jaime. They have endured with stoicism my dedication to the development of my academic courses and this thesis. For them, I can only say: "os quiero mucho, y muchas gracias por dejarme seguir en casa".

Second, I am in debt with my thesis advisor, Ana García Serrano, by her trust in this project, and the freedom and scientific creativity that have defined our relationship. I expect to keep her wise advise in our future academic challenges. She has proven to have an open mind to discuss and to accept new ideas, with novel points of view coming from fields, so far away at a first glance, like the geometry, the information retrieval and semantics.

By last, but not less important, I would like to express my sincerely gratitude to the teachers and advisors of the master, all of them, lecturers of the Language and Computer Systems of the UNED. It has been my first experience as student of the UNED, and I have found it a wonderful experience with regard to the means used, the content of the subject, the academic excellence in student's work plan and the evaluation method. All the subjects was oriented to the research, and the academic demand was quite high, nevertheless, the assignments were always very well oriented to gain a depth knowledge of some research topic. We got a few seed papers to start our search, and our devotion and freedom did the rest. In many cases, a devoted and careful reading led us to find some original gaps to be filled, such as the two research problems studied in this thesis, and an unpublished survey on sentiment analysis in Twitter.

Preface

"Three passions,
simple but overwhelmingly strong,
have governed my life: the longing for love,
the search for knowledge, and unbearable
pity for the suffering of mankind. "
Bertrand Rusell

The main objects of research in this thesis are two problems in information retrieval (IR). First, an ontology-based IR model for the indexing and retrieval of semantically annotated data, and second, a text classifier. Both problems are related by a common thread derived from a same geometric point of view, approach that proved to be helpful to find some gaps in the literature, derived from certain geometric inconsistencies in the current models. In both cases, the solution proposed to bridge the gap was also inspired by a geometry-based approach. We used basic results of well established geometric theories, such as differential geometry and the metric spaces.

The thesis is divided in two parts, one per problem. Historically, the first problem studied was the text classifier, called intrinsic Bayes-Voronoi, introduced in the second part of the thesis. We found some gaps in the standard VSM model, or "bag of words", during the spring of 2012, and we developed the whole theoretical model as an essay about text categorization for the subject about "Text mining", while the experiment validation was carried-out during the summer of this year, obtaining some discouraging results. As final essay of the same subject, we also carried-out an original survey on sentiment analysis in Twitter, which has been translated to the English, and coauthored with Ana García, but unfortunately, it has not published yet.

The conception of the idea about the ontology-based IR model, proposed in the first part of the thesis, emerged as the final essay of the subject about "Intelligent Information Retrieval", being submitted in September of 2013. In this preliminary essay we identified some gaps in the literature, which were derived from the geometric inconsistencies in the adaptation of VSM in some ontology-based IR models in the literature. Once the gap was identified, we proposed a novel IR model defined by a set of principled-based axioms whose main idea was to build a structure-preserving model, following the classical ideas about invariance, so popular in the field of applied geometry. During the academic course 2013-2014, we developed our ideas up to be able to propose an IR model that would fulfill these design axioms and could bridge the gap, arriving to the model introduced in the first part of the thesis, called *Intrinsic Ontological Spaces.*

The ideas about the preservation of geometric structures and transformation groups, back to the pioneer work of Felix Klein [Klein, 1893], also known as the *Erlangen program*. In his work, Klein introduces the concept of *group of transformations* and he defines as primary object of research for the geometry: the study of the invariant properties of the geometric objects under the action of the groups of transformations. Klein opens a conceptual revolution and change of paradigm in geometry, which affects the whole mathematics, whose more abstract descendant is the theory of categories introduced by Eilenberg and McLane in [Eilenberg & MacLane, 1945]. The last one is devoted to the study of the mappings of algebraic structures that preserve its intrinsic structures, also called natural equivalences. The ontology-based IR model proposed here is only a humble inquiry inspired by these theories.

By last, we expect to have introduced some novel ideas from a geometric point of view in the information retrieval field, whose reading can result helpful and interesting for any reader working in this exciting field.

Juan José Lastra Díaz
Madrid, 7th September 2014

Chapter 1

Introduction

The Vector Space Model (VSM) [Salton et al., 1975] is omnipresent in all sort of information retrieval systems, such as the classical keyword-based search engines, text categorization, question answering, and text summarization among others. In this work, we propose a couple of original semantic space models for two problems in information retrieval (IR): the development of a novel ontology-based IR model, and a keyword-based text categorization method.

In keyword-based text categorization, the state of the art has been defined, until recently, by the use of keyword-based systems based in the representation of documents as vectors in a VSM model, together with the use of SVM classifiers [Lewis et al., 2004]. The main drawback of the classical keyword-based IR models is their lack of meaning, which limit its search capabilities to objects mentioned in the text, without any possibility to infer any sort of relation between the concepts and entities in a query, and the keywords in the vocabulary of the model. Such as is noted in [Castells, 2008], the limitations of the meaningless VSM models have motivated the advent of a novel generation of ontology-based IR models, such as the pioneering works introduced in [Vallet et al., 2005] and [Fang et al., 2005].

Most of novel ontology-based IR models in the literature include any sort of adaptation of the standard VSM model with the aim to represent concepts or individuals within a populated ontology, like vectors in a vector space as means to can compare them. The novel models move from a keyword-based vector representation to a concept-based vector representation, wherein the base vectors of the model can be concepts or instances of concepts (entities, individuals).

1.1 Motivation

Our investigation is motivated by the identification of some geometric inconsistencies in the adaptations of VSM in some ontology-based IR models in the literature, or the classical keyword-based text classifiers as the described in [Lewis et al., 2004].

The ontology-based IR models use some adaptations of the VSM model to represent concepts, or instances of concepts, as base vectors in a VSM model, with the aim to be able to use standard IR techniques as the weighting and cosine-based ranking. However, these models overlook some important geometric properties and constraints, as well as its consequences. Among these overlooked concepts, we can

cite the orthogonality condition and the cardinality mismatch. In the case of classical keyword-based VSM models, we find a similar overlooked property that is the fact that the normalized vectors are defined on the unit hypersphere, not in the general Euclidean ambient space, thus the models should integrate this constraint in its mathematical representation. Precisely, the aim of IR models proposed in this thesis is to bridge this gap, as well as other drawbacks described throughout this work.

First, the current ontology-based IR models implicitly assume that we call the orthogonality condition, derived from the use of cosine function as similarity measure among vectors. Two different concepts, or instances of them, that share a common ancestor are represented by these models as independent base vectors, which are mutually orthogonal. This orthogonality condition means that two related concepts as *bicycle* and *motorbike* will have zero similarity, instead of a high value, such as we will expected for two concepts derived from a common concept called *two-wheel vehicle*. Other inconsistency is the cardinality mismatch, where objects with different cardinality are mixed in the same model. It is the case for the references to instances of concepts (individuals) or the mentions to whole classes (sets) which denote a collection of subsumed classes and individuals. In the introduction of the chapter 2 we provide an exhaustive description of many other drawbacks found in the literature.

In the case of text classifier based in VSM, the vectors are normalized to have unit norm. It means that all the vectors in a normalized VSM model are contained in the positive unit hypersphere, instead of the whole euclidean space, therefore, the normalized vectors have reduced their intrinsic dimension by 1. The text classifier introduced in chapter 3 is motivated by this gap, and our aim is to define a features space that represents the document vectors on the unit hypersphere, instead of the euclidean ambient space, such as is made by previous models.

Although, we only provide here a brief description about the motivation for the work carried-out, we explain our motivation with more detail in the introductory sections of the chapters 2 and 3, wherein we introduce the proposed models.

1.2 Research problems

The main research problems studied in this thesis are two as follows. First, the design of a novel ontology-based IR model (chapter 2) that bridges some gaps in the geometric structures of the models in the literature. Second, the design of a novel text classifier based in ideas from intrinsic geometry on differential manifolds (chapter 3).

In addition, we study the ontology-based semantic distances, such as the Jiang-Conrath distance [Jiang & Conrath, 1997], and we propose a novel generalization of the last one to obtain a well defined metric on any sort of taxonomy, unlike the standard JC distance, which is only a metric on tree-like taxonomies.

1.3 Contributions

This thesis introduces some novel contributions to the body of knowledge in the field of information retrieval (IR), such as follows:

1. A novel ontology-based IR model which preserves all the semantic structures encoded in any base ontology, called *Intrinsic Ontological Spaces*, which is pending to be validated experimentally.

2. A novel ontology-based semantic distance, called *weighted Jiang-Conrath distance,* which matches the standard definition of the Jiang-Conrath distance on tree-like taxonomies, while it generalizes its definition to guarantee that the novel distance is a well defined metric on any sort of taxonomy.

3. A novel ontology-based ranking based in the use of Hausdorff distance among subsets of a metric space, defined by the space of subsets of weighted-mention to instances and classes annotated within a populated ontology, which is used to represent semantically annotated data in the *Intrinsic Ontological Spaces.*

4. The definition of a metric based in the novel *weighted Jiang-Conrath distance*, which allows the integration of weighted-mentions to classes and individuals in a same semantic space, while it mimics the Jiang-Conrath distance among concepts on a tree-like taxonomy.

5. The definition of a family of novel semantic spaces, the *Intrinsic Ontological Spaces,* based in a metric space defined by a whole set of ontology-based components, such as the novel ontology-based semantic distance, a ontology-based weighting and ranking methods, and the integration of the classes and individuals in a same space, while the model preserves all their intrinsic semantic relations.

6. A novel text classifier, called *Intrinsic Bayes-Voronoi*, based in the use of a vector-based generative model with a normal distribution defined on the unit hypersphere.

7. The definition of a normal distribution on differential manifolds, called *geodesic Normal distribution*, and the *geodesic Mahalanobis distance* associated to the manifold-based distribution.

8. We also prove that any manifold-based normal distribution induces a Bayes classifier which is defined by a Voronoi diagram on the manifold.

1.4 Structure of the thesis

The thesis is structured in two independent chapters, one for each studied problem. The chapter 2 introduces a novel ontology-based IR model called *Intrinsic Ontology Spaces*, and a novel ontology-based semantic distance called *weighted Jiang-Conrath distance.* The chapter 3 introduces a novel text classifier called *Intrinsic Bayes-Voronoi*, and it also defines some novel statistical objects on differential manifolds,

such as the *geodesic Normal distribution* and the *geodesic Mahalanobis distance*. moreover, we prove that the Bayes classifier for any geodesic Normal distribution defines a Voronoi diagram on the manifold domain of the distribution. By last, the chapter 4 introduces a summary of our main conclusions and contributions, and in addition, it also introduces some research trends as future work, mainly the experimental validation of the novel ontology-based IR model.

The chapters 2 and 3 are self-contained, wherein every one fulfills the expected structure for any academic paper as follows: abstract, introduction, motivation, related work, preliminary concepts, description of the proposed models, experiments and results and by last, its conclusions and future work.

1.5 Publications

Our intention to apply for a patent has prevented any kind of dissemination of the content of this thesis, before the submission of the official patent application. For this reason, this thesis was defended in a private session on September 29, and it has not been disclosed for public dissemination until December 24, 2014.

The whole content of the chapter 2, composed by a novel ontology-based IR model proposed called *Intrinsic Ontological Spaces*, as well as a novel ontology-based semantic distance, called *weighted Jiang-Conrath,* has been submitted in the form of the patent application below [Lastra Díaz & García Serrano, 2014]. This publication also includes some novel edge-based intrinsic IC-computation methods not included herein.

By other hand, the Intrinsic Bayes-Voronoi text classifier has yielded some preliminary discouraging results, which has slowed our initial inclination to publish the preliminary results of the model.

By last, we developed a novel and exhaustive survey about the problem on sentiment analysis in Twitter, coauthored with Ana García Serrano, but unfortunately, it has not been published yet.

1. Lastra Díaz, J. J., & García Serrano, A. (2014). System and method for the indexing and retrieval of semantically annotated data using an ontology-based information retrieval model. United States Patent and Trademark Office (USPTO). US14/576,679. December, 19.

Chapter 2

A novel ontology-based IR model

This chapter introduces a novel ontology-based IR model called *Intrinsic Ontological Spaces*, and a novel ontology-based semantic distance called *weighted Jiang-Conrath distance.* The main idea of the model is to build an embedding of a populated ontology into a *metric space*, while its intrinsic geometry is preserved. The proposed model unifies the representation of the classes and individuals of the ontology in a same semantic space, while their intrinsic semantic structure relations are preserved. The documents, or any other sort of information units, are represented by sets of weighted-mentions to individuals and classes within the ontology, while the queries are represented as sets of mentions to individuals and whole classes, considering the last ones as sets of subsumed concepts and individuals. The representation space is defined by an extension of the *weighted Jiang-Conrath distance* among concepts, whose purpose is defining the distance among individuals, and a weighting scheme to represent documents and queries. The *weighted Jiang-Conrath distance* is defined as the shortest weighted-path among concepts, according to a generalization of Jiang-Conrath edge weights. Unlike the standard Jiang-Conrath distance, the *weighted Jiang-Conrath* is a well defined metric on any sort of taxonomy. The ranking method is based on a distance function among document and queries, which is defined by the Hausdorff distance among subsets on a metric space, according to the metric of the representation space. The proposed model is a well defined metric space on any ontology with a general poset structure. In the case of a tree-like base ontology, the representation space mimics exactly the Jiang-Conrath distance among concepts and individuals and it also verifies the structure of a hierarchical Voronoi diagram.

2.1 Introduction

The *Vector Space Model* (VSM) [Salton et al., 1975] is known as "bag of words", because every document is represented by a vector whose coordinates are defined as a function of the term occurrence frequency within a document. The set of terms used to represent every document is called the vocabulary of the model, and it defines the base vectors of the vector space. In most of cases, the cosine function is used as a similarity measure between a query vector and the vectors representing the indexed documents. Due to its simplicity, and the success achieved, the VSM model is the kernel for most of search engines, and it has been adopted in many tasks and applications of natural language processing (NLP), such as: information retrieval

(IR), document categorization (TC) and clustering, web mining and automatic text summarization (TS) among others.

Recently, the vector space models has been extended to define word and phrase spaces, such is reflected in [Erk, 2012], [Clark, 2012] and [Turney & Pantel, 2010]. A word or phrase space is a vector space where the vectors represent these information units instead of documents, and the space metric encodes the semantic similarity between pairs of information units. The word spaces are based in the distributional hypothesis [Basili & Pennacchiotti, 2010], which sets that words in similar contexts have similar meanings. In these models, the vectors representing every word are built as a function of the terms frequency in the context of one word within a document, so that these models allow encoding some semantic relations and statistics, such as the term cooccurrence, the synonymy and the meronymy among others.

Although the vector space models has been mainly used to represent text documents, such as we saw above, these models have been successfully applied to represent other types of information units, such as words, phrases and sentences. Following the previous reasoning, we state that the ontology-based IR model proposed here, works with any *information* unit that can be encoded in an ontology, according to the definition below.

The information units are the objects indexed by the ontology-based IR model proposed here, and it could be text documents, web pages, sentences, multimedia objects, or any sort of data that admit a ontology-based representation.

Definition 1 (Information unit) *An information unit is any sort of semantically annotated data that can be represented as a collection of concepts (classes) or instance of them (individuals) within an ontology.*

The main limitation of the VSM model is its lack of meaning. As is noted in [Castells, 2008], most of the current web search systems use a standard VSM model with meaningless terms, which make impossible to retrieve documents using queries with non-explicitly mentioned terms in the corpus. By other hand, we can appreciate the same situation in other related problems where the same meaningless version of the VSM model is used, such as in the text categorization problem [Sebastiani, 2002a], [Lewis et al., 2004].

The advent of the *semantic web* has motivated a great change of paradigm in the IR community. The IR models has moved from a model based in meaningless terms to a model based on references to concepts or its instances, namely, there has been a change from a keyword-based paradigm to another concept-based one. The novel paradigm has converted the conceptual models and the knowledge bases in its core components, and ontology languages, such as OWL, have become the favorite representation to encode this knowledge and to store the references to the indexed data. Nowadays, the use of ontologies is omnipresent in all sort of semantic retrieval tasks in the context of semantic web [Ding et al., 2007], as well as in other application fields like the bioinformatics [Pesquita et al., 2009].

Motivated by the lack of meaning in previous IR models, some novel conceptual IR models have appeared during the last decade, whose main example is the family of *ontology-based IR models*, whose abstract definition is given below.

Definition 2 (Ontology-based IR model) *An ontology-based IR model is any sort of information retrieval model which uses an ontology-based conceptual representation for the content of any sort of information unit, whose main goal is the indexing, retrieval and ranking regarding of these infomration units with regard to any user's query.*

What is the main contribution provided by the use of ontologies in IR models ?. The essential contribution of the ontologies in IR is the capability to retrieve documents semantically related to concepts or entities not mentioned in the query or any document. From an abstract point of view, the ontology-based models make a virtual expansion of the semantic objects (concepts/instances) in the queries and the documents before to compare them. These models allow to retrieve documents using concepts or entities not directly mentioned in these documents.

We subdivide this family of ontology-based IR models in two subfamilies: (1) the *vector ontology-based IR models*, whose main feature is the use of some adaptation of the standard VSM model to manage concepts instead of meaningless terms, and (2) the *ontology-based metric space IR models*, whose unique examples, up to our knowledge, are the pioneering work of [Rada et al., 1989] and this work. Among the main works in the subfamily of vector ontology-based IR models we can cite the pioneering works in [Vallet et al., 2005], [Fang et al., 2005], [Castells et al., 2007], [Mustafa et al., 2008], [Dragoni et al., 2010] and [Egozi et al., 2011] among others.

By other hand, the unique ontology-based IR model based in a metric space that we have found, up to our knowledge, is the model proposed in [Rada et al., 1989]. This work introduces some ideas that are closely related to the model proposed here, although we can also find some important differences that will be explained in our review of the state of the art.

To the best to our knowledge, the work in [Rada et al., 1989] can be considered as the oldest reference within the ontology-based IR family. Surprisingly, this work is not cited by others ontology-based IR models found in the literature, despite that the Rada's measure is highly cited and well known in the scope of the ontology-based semantic distances. Roughly speaking, most of the ontology-based IR models reported in the literature models, such as the cited in this work, do not have taken advantage of the results in the field ontology-based semantic measures, precisely, unlike our work.

From an abstract point of view, the main features of the family of vector ontology-based IR models, also called adapted-VSM models, are as follows: (1) the use of a conceptual representation for documents and queries based in an ontology, (2) the retrieval of relevant documents through any ontology query language, (3) some sort of vector space for the representation of references to concepts and instances, based in a set of orthogonal base vectors defined by the classes and individuals of the ontology, (4) some sort of adaptation of standard term-frequency weights for the definition of coordinates, (5) the use of the cosine function as ranking method to sort the relevant documents, and (6) a multivector representation and ranking combining different types of features, such as concepts, keywords or ontological features.

This work introduces a novel structure-preserving ontology-based IR model, called *Intrinsic Ontological Spaces*, for the indexing and retrieval of semantically annotated data, such as text documents, web pages, or any sort of information that

can be represented as a set of semantic annotations (individuals and classes) on any sort of base ontology. The proposed model bridges the gap of geometric inconsistencies in current methods, such as is explained below. The work includes a detailed description and justification of the model, and a very simple example to explain its operation. Nevertheless, it still be pending its experimental validation. The main expected benefits of the proposed model are an improvement in terms of the ranking quality, as well as in the precision and recall measures. Our hypothesis is that our model should contribute to the joint improvement of these evaluation measures, thus, it should contribute to improve the results expected by the users of any information search system based in the proposed IR model.

2.1.1 Motivation

A vector space is a very rich and versatile algebraic structure that, precisely by its versatility, has been used in an irreflexive manner in the field of information retrieval. Formally, a vector space is an additive Abelian group with a scalar product that is associative and distributive, it means that the space vector includes all the inverse elements for each document, and every linear combination among them; nevertheless, all these elements of the space are not used, or required, in any IR model. Maybe, the main reason to use vector spaces in the current IR models is to rank the documents using the cosine function as similarity measure, due to its simplicity, computational efficiency and the success achieved in many tasks in information retrieval.

The state of the art in ontology-based IR models has proven the potential benefits derived from the use of conceptual models with regard to the meaningless IR models. However, if we study carefully some overlooked assumptions in these conceptual models, we find some important aspects that offer an important improvement opportunity in terms of ranking quality, as well as in the precision and recall measures.

Main motivation behind most of the adapted-VSM models have been to build a semantic weighting method to compare semantically annotated documents, however, these models have been using the vector space model as a black-box without take into account some important implicit assumptions of the model and its consequences. Making a review of the current literature about the topic, we find the following gap which motivates our work:

Orthogonality condition. The base vectors of any VSM model are mutually orthogonal, it means that similarity cosine function among different base vectors is zero. One consequence of the orthogonality condition of the adapted VSM models is that two vectors associated with two documents can get a zero, or very low similarity value, when they do not share references to the same concept instances, although these instances could share a common ancestor concept in the taxonomy. By example, documents with references to bicycle and motorbike models would not be related, although the instances are derived from the two-wheel vehicle concept.

Cardinality mismatch. Some of these ontology-based models are not including references to classes as sets of the objects, or they are mixing references to classes and instances (individuals) at the same representation level. The main

idea behind most adaptations of the VSM models to manage the ontology information is to make a mapping from individuals and/or classes to base vectors of the representation vector space. In this way, the models are assigning two different, and opposite meanings, to the same base vector. In one case, the base vector represents the occurrence of one object (individual), otherwise, a base vector is representing a collection of objects (classes). These inconsistencies can be summarized as a cardinality mismatch in the adapted VSM models, and the nature of the objects represented by the model.

Statistical fingerprint vs. semantic distances. The metric used to compare documents by most of published ontology-based models is based in the Euclidean angle among normalized vectors (cosine score). The vectors encode the statistical fingerprint of the indexed documents, it means, the statistical cooccurrence relations among different concepts in a document, but this metric lacks of a meaning in the sense that they are not encoding any semantic distance among concepts, such as it is made by very well established ontology-based distances, such as the Jiang-Conrath measure [Jiang & Conrath, 1997]. The only exception to the problem described here is the IR model proposed in [Rada et al., 1989], which defines a Boolean semantic model, where the documents are represented by sets of concepts, although the concepts are annotated in binary form without using any semantic weighting method, such as is provided by our method. The last model is closely related to our work and we consider this model, up to our knowledge, the first published ontology-based IR model.

Populated ontology are not directly indexed. Many of the ontology-based IR VSM models need to retrieve the related documents with the instances and concepts in the query before to rank them. The populated ontology is not indexed directly, by this reason, it needs to be searched using any ontology-based query language, such as SPARQL or any other. By contrast, our model builds a direct geometric representation of the data in the populated ontology, integrating retrieval and ranking in a same step. Despite this approach could produce bottlenecks for large scale ontologies, we expect that the integration of geometrical search structures in the model allows to speed-up the queries.

Lack of a semantic weighting. The weights in adapted VSM models are statistical values, not related to the real semantic weight of the concept/instances within the document.

Continuity problems on metrics on sets. In [Rada et al., 1989], the authors introduce an ontology-based IR model which defines a metric space using a shortest path metric on the taxonomy, and the average distance among sets of concepts as distance function among documents. The authors report some continuity problems around close documents, The source of the problem is that their distance function among documents does not fulfill the coincidence axiom of a metric (see definition 6), thus it is not a well defined metric on sets, unlike the Hausdorff distance used in our model.

The Jiang-Conrath distance is not well defined. Some recent research has unveiled that the Jiang-Conrath distance only satisfies the metric axioms for tree-like ontologies [Orum & Joslyn, 2009]. This fact contradicts the original statement of the authors in [Jiang & Conrath, 1997]. The Jiang-Conrath distance depends on the lowest common ancestor between two concepts, which is only uniquely defined for lattices, not for general posets. Despite of the JC distance is well defined on lattices, in [Orum & Joslyn, 2009] the authors provide some counterexamples to demonstrate that neither in this case the JC distance is a metric. We also provide a counterexample to enlighten the problem in figure 2.11. In section 2.4.4, we introduce a generalization of this measure to fulfil the metric axioms on any sort of taxonomy.

2.1.2 Research problem and main hypothesis

The main goal of the work in this chapter is to propose a novel structure-preserving ontology-based IR model to bridge the gap described above. Our work follows a geometric approach in the definition of the gap to be filled, as well as in the proposed solution. We propose as solution to bridge the gap, a novel ontology-based IR model called *Intrinsic Ontological Spaces*, which is based in the integration of the intrinsic structure of the base ontology in the definition of the representation space itself, our approach can be interpreted as a *semantic metrization* of the populated ontologies.

The main hypothesis behind our IR model is that the integration of the structures encoded within the base ontology of the IR model in the representation space must contribute to the improvement of the ranking quality, and the precision and recall measures expected of the proposed model, with regard to prior models.

2.1.3 Summary of the chapter

This chapter introduces a novel ontology-based IR model called *Intrinsic Ontological Spaces*, as well as a novel ontology-based semantic distance called *weighted Jiang-Conrath distance*. The purpose of the method is to provide an ontology-based IR model for the indexing and retrieval of semantically annotated data, such as text documents, web pages, or any other sort of information units that can be represented by semantic annotations within a base ontology.

The main idea of our model is to build a structure-preserving embedding of a populated ontology into a metric space, while its intrinsic geometry is preserved. The proposed approach fills a gap of modelling inconsistencies in current methods, whose consequence is a best ranking, precision and recall measures. The proposed IR model unifies the representation of the classes and individuals of the ontology in a same semantic space, while their intrinsic semantic structure relations are preserved. The text documents, or any other sort of semantically annotated units, are represented by sets of weighted-mentions to individuals and classes in the ontology, while the queries are represented as sets of mentions to individuals and classes considered as sets of subsumed concepts and individuals. The representation space is a metric space defined by an extension of our novel weighted Jiang-Conrath distance among concepts, whose purpose is defining the distance among weighted mentions to individuals and classes, and a weighting scheme to represent documents and queries.

The intrinsic geometry of any ontology is defined by three algebraic structures: (1) the order relation of the taxonomy, (2) the intrinsic semantic distance among the classes and individuals, and (3) the set inclusion for the individuals and subsumed classes of the ontology.

The model proposed in this work comprises the following elements: (1) the definition of the semantic representation space as the universal set of weighted-mentions to individuals and classes within the populated base ontology, that we call Intrinsic Ontology Spaces; (2) an embedding method to injects semantically annotated data, or information units, in the representation space of the model; (3) an embedding method to injects semantically annotated queries in the semantic representation space of the model; (4) a semantic weighting method that combines statistical and semantic information to represent the semantic annotations associated to the indexed information units in the semantic representation space; (5) a novel ontology-based semantic distance among concepts (classes) and instances of concepts (individuals) within a populated base ontology, that we call weighted Jiang-Conrath distance; (6) a novel ontology-based ranking method for the retrieval and sorting of the indexed units retrieved by the system, and (7) a pre-processing step whose purpose is computing all the parameters and data structures to enable the indexing and searching operations of the search engine of the system.

The representation space is a metric space defined by a generalization of the Jiang-Conrath distance among concepts [Jiang & Conrath, 1997], which we call *weighted Jiang-Conrath distance*. The purpose of this novel ontology-based distance is integrating the individuals in the representation, defining a metric on any sort of ontology, and allowing the definition of a semantic weighting scheme to represent documents and queries, while it overcomes the drawbacks of the standard Jiang-Conrath distance.

The weighted Jiang-Conrath distance is defined as the shortest weighted-path metric among concepts and individuals within the base ontology, according to the generalized Jiang-Conrath edge weights (IC difference). The weighted Jiang-Conrath distance is a well defined metric on any sort of taxonomy. By contrast, the Jiang-Conrath distance only is a metric on tree-like ontologies [Orum & Joslyn, 2009]. The proposed model could use any known intrinsic IC-based method to compute the IC-values for every concept in the base ontology, although our preferred approach is the method proposed in [Pirró & Seco, 2008]. The IC-values depends only on the structure of the ontology, thus, these can be computed a priori during the set up process of the search engine. By other hand, the computation of the shortest weighted-path among concepts needs to be done through any Djikstra-type algorithm [Ahuja et al., 1990], with the inconvenient that it could be very expensive for large base ontologies. By this reason, we define a pre-processing step to compute a priori the IC-values and all the pairwise distances among concepts of the base ontology, such as is shown in figure 2.13.

The ontology-based ranking method proposed in this work is based on a distance function among document and queries, which is defined by the Hausdorff distance among subsets of a metric space, according to the semantic metric of the representation space. The use of the Hausdorff distance and the weighted Jiang-Conrath distance guarantee that the model is a well defined metric space on the space of all annotated information units (documents) on any sort of base ontology.

Up to our knowledge, the IR model proposed in this work is the first one to use an ontology-based semantic distance and the Hausdorff distance as ranking method, unifying the representation of weighted-mentions to classes and individuals in a same metric space. Moreover, in the case of a tree-like base ontology, the representation space also verifies the structure of a hierarchical Voronoi diagram [Gold & Angel, 2006], where every parent concept geometrically subsumes its descendant concepts. We could say that the logic hierarchy of the base ontology is transformed in a geometric hierarchy according to the semantic metric of the model.

The proposed model unifies the representation of the classes and individuals of the ontology in a same semantic metric space, while their intrinsic semantic relations are preserved. The documents, or other information units, are represented by sets of weighted-mentions to individuals and classes in the ontology, while the queries are represented as sets of mentions to individuals and classes considered as sets of subsumed concepts and individuals.

Our model avoids the use of vector spaces to rank documents, instead, a document (information unit) is defined as a collection of weighted mentions to classes and individuals, and the ranking method for documents is built using the metric of the space and the Hausdorff distance among subsets. The mentions to classes (concepts) within a user query are mapped to subsets of the representation space, while the mentions to classes in the documents are managed as weighted mentions to distinguished individuals of the parent class.

Structure-preserving ontology embedding. The main feature of the proposed model is that the embedding of the information units in the semantic representation space preserves the three main structure relations of the ontology defined as follows: (1) the intrinsic semantic distance (metric structure) among classes, (2) the taxonomic relations (topological/order structure), and (3) the set inclusion relations (set structure). We call these three structure relations as the ontology intrinsic structure.

We can say that the proposed model builds a natural equivalence between the information units and its embedding in the representation space, following the notion of natural morphism in the theory of categories [Eilenberg & MacLane, 1945]. The proposed model tries to capture and to save all the semantic information provided by the ontology, avoiding any information lost in the embedding process. The input for our ontological space is a populated ontology with semantic annotations of any sort of information units. It means that the model assumes the existence of a complementary semantic annotation module whose aim is to search the references to classes and entities of the ontology.

The *Intrinsic Ontological Spaces* model allows ranking documents, or any other semantically annotated data, using a semantic distance function derived from the ontology model. We expect that our model allows to improve the ranking quality and the precision and recall measures of the current methods, while it solves the inconsistencies in the current models.

The proposed solution has some theoretical and practical advantages over current methods:

1. The proposed IR model removes some inconsistencies in previous models, such as the: orthogonality property and the cardinality mismatch already explained

above; the lack of a ranking method based in an intrinsic semantic; the lack of a semantic weighting method. The removal of these inconsistencies contributes to get an improved semantic representation model, whose main consequence is the improvement of the ranking quality and the precision and measures for any application based on the novel IR model.

2. All the logic components of the IR model are ontology-based. It includes the retrieval process, the weighting schema, the ranking and the definition of the representation space itself. Every element of the IR model is directly derived from the structure relations encoded by the base ontology used for the indexing of the data.

3. The proposed IR model allows integrating many geometry-based algorithms and theoretical results with potential benefits for the model. For example, we can integrate well known geometry-based space search methods to find nearby documents [Brin, 1995], enabling the extension of the model to large scale document collections as the web, or large text repositories in government agencies and private companies.

4. The ranking and weighting computation model that is proposed can be estimated on-the-fly without any training phase, because our weighting method does not use an inverse frequency table.

5. The factorization of the weights for the mentions in static (normalized frequency of the mentions) and dynamic (IC-value per concept) factors, allow updating the input parameters of the model (IC-values) while the index form of the indexed units is preserved. It means that the ontology could be dynamically updated in different ways (merge, concept insertion, etc.), without to make changes to the indexed units, while the parent classes for the weighted-mentions still are in the ontology. By example, if a set of new classes is added to the base ontology, the system only needs to run the pre-processing step to get the new set of distances among concepts (classes), then, any new query answer will be computed using the novel semantic relations in the base ontology.

6. Like other known methods, our model also merges the retrieval and ranking of documents in a same step, removing in this way the necessity to use SPARQL or any other query language to retrieve the documents to be ranked, as well as any other semantic retrieval method as is proposed in [Mustafa et al., 2008]. The query is represented as any another document, and it is used to search the full set of indexed documents, eliminating the first retrieval step of the current models.

Throughout our discussion, we use a language inspired in geometric notions whose purpose is to enlighten some analogies and relations between the conceptual spaces and its geometric images, expecting it allows us to use all these well established and powerful theories. Our work is related to some results in a novel research trend called geometry-based semantic models, whose first reference is defined by

some preliminary ideas discussed by Widdows [Widdows, 2004] and partially developed by Clarke in his thesis [Clarke, 2007].

As we discussed above, our own work is a novel contribution to the family of ontology-based IR models, and to the family of ontology-based semantic distances. By last, the Intrinsic Ontological Spaces can also be interpreted as a extreme way of ontology-based query expansion, problem recently revised in [Wu et al., 2011], where all the admissible query expansions are already integrated in the representation space itself, avoiding the query expansion problem. In this last direction, in [Saruladha et al., 2012] the authors studied how the semantic similarity/distance measures could be used to expand the user's query through the use of semantically related words in an ontology, problem that is avoided by our model.

2.1.4 Potential applications

As we said above, we already mentioned that the use of any sort of vector space models is omnipresent in all sort of natural language applications, specially as IR models for all sort of web and data search engines. The ontology-based IR model proposed in this work defines a new paradigm for the semantic indexing of all sort of semantically annotated data, whose main goal is transforming the search processes made by the users from a keyword-based search to a concept-based search. Therefore, our IR model can be considered as a complement and a new generation of IR models destined to substitute the current generation of keyword-based search engines, including also the recent ontology-based IR models based in adaptations of the VSM model.

The proposed model is framed in the family of ontology-based IR models, and it shares a common goal with other previous methods: be the cornerstone of a new generation of semantic search systems.

As the VSM models, the proposed model can be applied in the context of any NLP application where any sort of semantic space is used, among we can cite: web search system, any sort of IR system for text indexing and retrieval, cross-language information retrieval (CLIR) systems, automatic text summarization systems, text categorization and clustering, question answering systems, and word disambiguation among others. Moreover, the proposed model also can be applied in bioengineering applications where the data and the domain knowledge are represented within a domain-oriented ontology.

The ontology-based IR model proposed in this work is able to update any sort of application based in semantic vector spaces, or ontology-based adapted VSM models. By example, the VSM model has been extensively used in the context of cross-language IR systems (CLIR), such as the model shown in figure 2.1.

Other problem where some adaptations of VSM have been proven its utility, and our model could be applied, is in automatic text summarization (TS). In the scope of extractive TS methods we can find some conceptual models based in adaptations of the VSM model to represent the semantic similarity relations among sentences. This is the followed approach in [Meng et al., 2005] where the authors define the conceptual vector space model (CVSM), whose ideas are very close to the ontology-based IR model approach. Other TS methods are based in the clustering of sentences, originating the notion of centrality, whose core idea is that any document can be represen-

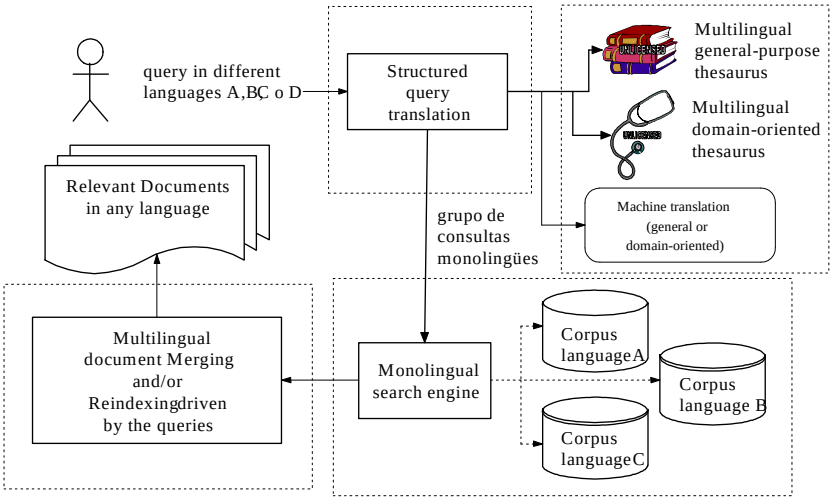


Figure 2.1: Basic architecture of a MLIR/CLIR system

ted by the more significative (central) sentence. These clustering methods use a VSM model to represent the sentences in any document, where each vector encodes a set of features of the sentence, and the model can use different functions to establish the similarity among sentences. Among these clustering TS methods we find the pioneering works of [McKeown et al., 1999] and [Hatzivassiloglou et al., 2001], as well as the works in [Siddharthan et al., 2004] and [García-Hernández & Ledeneva, 2009]. Finally, the most recent text summarization (TS) methods are based in graph-ranking algorithms derived from PageRank and HITS, whose main references are the works of [Erkan & Radev, 2004], [Mihalcea & Tarau, 2004], [Wolf & Gibson, 2004] and [Vanderwende et al., 2004]. If the sentences within a document are considered as information units, these graph-based methods could benefit from the proposed model in this work, because the graphs are derived from the semantic similarity among sentences obtained through adaptations of a vector space model and a set of semantic features.

In the scope of the Q&A systems, the vector space models have been used to represent sentences within a document, and to retrieve text fragments with potential answers to a question. These approaches are inspired in IR models, and jointly with other techniques, they have been successfully proven in DeepQA [Chu-Carroll et al., 2012] to retrieve relevant text fragments for a user query.

Finally, other potential application of the Intrinsic Ontological Spaces is the word disambiguation problem, where have been proposed methods based in the vector representation of the context of a word [Navigli, 2009], following the distributional hypothesis. Due to the omnipresence of the vector space models in NLP, it is very clear that the proposed model has many potential applications in the scope of NLP and IR applications.

2.1.5 Structure of the chapter

The remainder of the chapter is structured as follows. In section 2.2 we make a survey of the state of the art in ontology-based IR models and geometry-based ontology embeddings. Due to its extension, we divide the state of the art in three parts:

(1) ontology-based IR models, (2) geometric representations for ontologies, and (3) ontology-based semantic distances. In section 2.2.3, we make an introduction and review of the semantic distances and similarity functions defined among concepts in an ontology, ideas that are a core element in the definition of our model. For sake of completeness, the section 2.3 makes an introduction of the basic concepts used in our discussion, such as ontologies, lattices, metric spaces, metrics on lattices and the Hausdorff distance, although any reader familiar with these concepts can skip the section. In section 2.2.3.2 we summarize some facts about the Jiang-Conrath distance in the context of our investigation. The section 2.4 introduces our IR model, there, we make a detailed description of the main components of the *Intrinsic Ontological Space* model: the indexation, retrieval and ranking methods. In section 2.5, we introduce a toy example to show the proposed model, and we briefly discuss the future experiments that should be carried-out for the validation of the model. By last, we elaborate and describe our conclusions and future work.

2.2 Related work

Our model is mainly related with three categories of works in the literature: (1) the ontology-based IR models, (2) the geometric representations for taxonomies, and (3) the ontology-based semantic distances.

According to the research problem studied, our work is framed in the family of ontology-based IR models, while that according to the approach followed in the proposed solution, our work is strongly related with the family of ontology-based semantic distance and similarity measures. In fact, our model includes a novel ontology-based semantic distance called weighted Jiang-Conrath distance. By other hand, the geometric approach adopted in our work is inspired by the same geometric spirit in the pioneering works about geometry and meaning of [Widdows, 2004] and [Clarke, 2007].

The remainder of the section is structured as follows. First, we review the state of the art about the ontology-based IR models. Second, we review some methods and ideas for the geometric representation of taxonomies that are related to the core ideas of the IR model proposed in this work. By last, we introduce and review the state of the art about ontology-based semantic distances, and we enumerate the known facts and drawbacks about the Jiang-Conrath distance, which have motivated the development of the novel semantic distance that we call weighted Jiang-Conrath distance.

2.2.1 Ontology-based IR models

Browsing the literature, we find some previous surveys about the family of ontology-based IR models. By example, we can cite the reviews made in [Castells, 2008] and [Fernández et al., 2011], as well as the survey in the context of multimedia retrieval made in [Kannan et al., 2012]. In other work [Wu et al., 2011], the authors survey the query expansion problem in IR and others ontology-based IR models, such as the analyzed ones in this section. Although our analysis of the state of art is exhaustive, the surveys cited can be useful to the reader to follow our analysis of the state of

the art about the main research problem studied here.

Ontology-based IR models and query expansion. Having cited the query expansion problem, reviewed in [Wu et al., 2011], we would like to express our view about the relation between this approach and the ontology-based IR models that we studied here. *We could consider the query expansion approach the dual of the ontology-based models*: the first one expands the query, while the last one expands the conceptual representation of the document. In the query expansion approach, the terms in the query are expanded with synonyms, related concepts or semantic annotations, and the expanded vector of terms is used to interrogate an unstructured semantic representation space. By other hand, in the ontology-based approach, the representation space is structured, and the semantic relations are already implicit in the indexation model, therefore the semantic representation of the documents is already expanded to match the queries in its base form.

In [Castells, 2008], the author makes a literature survey about the use of ontologies in IR and web mining, approach commonly known as semantic web, while he describes his experience in the development of an ontology-based IR system introduced in [Castells et al., 2007]. In other recent work [Fernández et al., 2011], the same group of authors introduce some extensions to the model in [Castells et al., 2007], in order to operate at web scale, while they also extend their previous literature survey. In [Kannan et al., 2012], the authors survey the ontology-based IR models in the context of the multimedia IR field.

For sake of understanding of the literature about the topic, the reader can see, in the tables 2.1 and 2.2, a summary of the main features of the ontology-based IR models analyzed in this section. The ontology-based IR models have been categorized in two subfamilies according to the structure of its representation space: (1) metric-space models, like ours, and (2) adapted VSM models.

Ontology-based metric space models					
IR Model	Doc. Rep.	Doc. Space	Retrieval	Weighting	Ranking
Rada et al., 1989	Set of boolean concepts	Ontology-based metric space (shortest path)	Integrated in ranking	Boolean	Average distance among sets of concepts
This work, 2014	Set of weigthed instances and concepts	Intrinsic ontology-based metric space (extension of JC distance)	Integrated in ranking	IC-based TF weights	Haussdorff distance among sets of weighted concepts and instances

Table 2.1: Ontology-based IR models based in semantic metric spaces

2.2.1.1 Metric space models

As we mentioned in the introduction, and to the best of our knowledge, the first published ontology-based IR model is proposed in [Rada et al., 1989]. The main

motivation of this work is the development of a IR model for biomedical applications, where the documents are represented as sets of concepts within a common ontology.

In [Rada et al., 1989], the authors propose to use the shortest path between concepts on an ontology as a measure of its semantic distance, and they call this measure *Distance*. The proposed IR model represents the documents and the queries by the set of concepts referenced in these information sources, nevertheless, the proposed IR model lacks of any weighting method, being a Boolean model. The documents are represented by the concepts associated to the instances in the document, but unlike our model, the instances are not represented in the model. To rank the documents according to a user query, they extend the *Distance* function among concepts to sets of concepts to define in this way a distance measure among documents. The *Distance* between sets of concepts (documents) is defined as follows: given two documents or queries, its ranking distance is defined as the average distance among all the pairwise combinations of concepts in the two sets.

Following the review of the model in [Rada et al., 1989], we find other drawback related to the continuity of their distance function. They define the distance among documents like the average pairwise distance among concepts in opposite sets. The authors reports an undesired continuity problem near the zero distance value. The source of the problem is that their distance function among documents does not fulfill the *coincidence axiom* of a metric (see definition 6), thus it is not a well defined metric on sets, such as the Hausdorff distance used in our model. With the aim that the distance function on sets of concepts can satisfy all the axioms for a metric, the distance function among sets of concepts is artificially forced to be zero when the two input sets of concepts are equal. A closer look to their formula in [Rada et al., 1989, def. 2, pp. 22] unveils an important underlying difference with the classical Hausdorff distance: while the Hausdorff distance uses the minimum-based point-set distance (definition 8), the implicit point-set distance in the Rada's model is the average distance among every point and all the elements in the set. This difference prevents that their distance function be a well defined metric.

The IR model proposed in [Rada et al., 1989] is very close in spirit to the IR model proposed in this work. We can find some similarities and differences among both models in some aspects.

First, we find that both models share some features: (1) both models use some sort of ontology-based semantic distance, (2) they representate the documents by sets (not vectors) of concepts, and (3) both share the definition of a rank function among sets of concepts.

By other hand, we find some differences as follows. First, both models represent documents by sets of concepts, although the *Intrinsic Ontological Spaces* also includes instances of concepts (individuals). Second, both models use a semantic distance defined on the ontology, but while Rada et al. use the shortest path length, we use a generalization of the Jiang-Conrath distance [Jiang & Conrath, 1997], which was designed to remove some known drawbacks in the edge-counting family of semantic distances and the standard Jiang-Conrath distance, such as we explain in depth in section 2.2.3.1. Third, the Rada's model use the average distance among all cross-pairs of elements to define a metric among sets of concepts, while we use the standard Hausdorff distance as metric with the advantage that the Hausdorff distance is well founded from a mathematical point of view. The Hausdorff distance

is the a metric on subsets of a metric space which is derived an extension of the metric of the space to sets. The Hausdorff distance is always continuous according to the topology induced by the metric of the space, removing the drawback related to the continuity around zero that Rada et al. report for their ranking function. It is interesting note that the ranking function proposed in [Rada et al., 1989] is very close to the definition of the Hausdorff distance, with the difference that the Hausdorff distance selects the maximum distance among all the point-set distance values, instead of the average value proposed by their model.

2.2.1.2 Adapted and enriched VSM models

The more recent family of ontology-based IR models start with the pioneering works in [Vallet et al., 2005] and [Fang et al., 2005]. Both works were independently published in very close dates, without any cross citation between them, or in others subsequent works as [Castells et al., 2007] and [Fernández Sánchez, 2009]. The IR model proposed in [Vallet et al., 2005] was continued in [Castells et al., 2007], being this research trend the core of the thesis of Fernández [Fernández Sánchez, 2009].

In [Vallet et al., 2005] and [Castells et al., 2007], the authors propose an ontology-based IR model based in one adaptation of VSM to represent concepts and individuals instead of meaningless terms. This model includes most part of the features exhibited by the models in the ontology-based IR family, and it could be considered as its canonical representative.

The main idea in [Castells et al., 2007] is to substitute the keywords vocabulary of a classic VSM, which defines the base vector set, by a vocabulary of *concepts* and *instances* within the base ontology of the KB, instead of a collection of meaningless terms. The documents are represented (indexed) by a vector of adapted TFIDF weights, where each weight is defined according to the saliency of a concept or instance of a concept within a document, and its *semantic discrimination* capacity. Each document is represented by a *set of concepts* and *concept instances*, instead of keywords, in this way, the system index the documents using concepts and instances as base vectors of its VSM model. To index the documents, the system associates a set of *semantic annotations* for the found references to concepts in the KB, which define the *collection of concepts* instantiated within each document.

The automatic semantic annotation is a very complex task which still be a very active research field in the information extraction (IE) community, by this reason, we consider the automatic semantic annotation problem out of the scope of our investigation, such as is made in [Castells et al., 2007], and we assume that the IR model proposed here need to be integrated with additional IE components for this task.

The operation of the IR model proposed in [Castells et al., 2007] is as follows. First, the system only accepts user queries in SPARQL format and it assumes that the documents have already been semantically annotated. Second, each document is represented by a set of semantic annotations in an ontology, which are defined by the references to concepts found in the documents. Third, the SPARQL query is used to interrogate the ontology and to retrieve all the documents with annotations derived from the concepts and instances included in the query. Fourth, all the documents retrieved are represented by vectors before to be ranked, while the base

of the vector space is defined by all the concepts and instances (individuals) included in the ontology, and an adaptation of TFIDF weighting scheme is used to convert the set of annotations of each document in a normalized vector expressed in the base of the concept vector space. By last, the retrieved documents are ranked using the cosine function.

The system, such as is expressed by the title of the work, is a direct and natural adaptation of the classic VSM model to manage concepts. The proposal in [Castells et al., 2007] agrees with other cited authors in that the proposed semantic IR model needs to be combined with standard keywords-based VSM models, due to the impossibility to have wide covering ontologies in a near future, by this reason, their system builds two independent VSM models (keywords + concepts) that are combined in the last retrieval stage.

The semantic retrieval capability of the Castells-Fernández-Vallet model is derived from the semantic retrieval of annotated document in the ontology, which is able to retrieve documents with references to concepts not included in the query or the document, starting from more abstract concepts defined in the base ontology. This capability is the essential contribution provided by the use of ontologies in IR, as well as the main reason for its broad acceptation in all sort of semantic search applications.

The documents retrieved by the Castells-Fernández-Vallet model are the documents annotated with entities found in the document collection retrieved by the SPARQL query [Castells, 2013], but the work do not clarify how it manages, if it does, the references to classes of concepts, it means when any document cites a set of objects using the name of the class, not a specific instance.

The model in [Castells et al., 2007] was extended in [Fernández Sánchez, 2009] and [Fernández et al., 2011] for broadening its application to a large scale and heterogeneous context as the web. Meanwhile, in [Bratsas et al., 2007], the authors introduce an application of the model in [Castells et al., 2007] to the problem of information retrieval in biomedicine, using a domain specific ontology and a fuzzy query expansion.

In [Fang et al., 2005], the authors propose an ontology-based IR model almost identical to the model in [Castells et al., 2007]. The model of Fang et al. has the same functional structure that the Castells-Fernández-Vallet model. The system admits queries defined by keywords or complex expressions which are transformed to queries in format OWL-DL. The OWL queries retrieve the related RDF triplets contained in the KB with references to the concepts and instances included in the user query. From the concepts and instances in the RDF triplets, the system retrieves the associated documents, and by last, the documents are ranked according to the user query. Such as in [Vallet et al., 2005], the model in [Fang et al., 2005] builds an adapted VSM representation trough a TDIDF weighting scheme using the instances-document frequency matrix, but unlike [Vallet et al., 2005], the final weights include a saliency factor whose purpose is to take into account the semantic differences among concepts and instances.

Concept-based adapted VSM models					
IR Model	Doc. Rep.	Doc. Space	Retrieval	Weighting	Ranking
Fang et al., 2005	Weighted bivector of instances and keywords	Instance–keyword based bivector VSM	OWL-DL queries	TFIDF + semantic saliency factor	Bivector cosine score combination
Vallet et al., 2005 Castells et al., 2007	Weighted bivector of instances and keywords	Instance–keyword based bivector VSM	SPARQL queries	TFIDF	Bivector cosine score combination
Mustafa et al., 2008	Concept-based weighted vector	Concept-based single vector VSM	Pairwise semantic distance among sets of concepts (query and document)	TFIDF	Combined score (cosine function + edge-counting semantic distance)
Dragoni et al., 2010	Wordnet concept-based vector	Concept-based single vector VSM	Integrated in ranking	TFIDF	Cosine score
Egozi et al., 2011	Vector of weigthed keywords + Wikipedia ESA-concepts	enriched ESA concept-based single vector VSM	ESA-based retrieval (doc-passages ranking)	Combined ESA-concept cosine score for passages and full document	Combined concept-based cosine score document-passages vs query
Cao & Ngo, 2012	Multi-vectors of weighted keywords + ontological features	enriched concept-based multi-vector VSM	Integrated in ranking	TFIDF	Barycentric combination of multiple cosine scores
Machhour & Kassou, 2013	Vector of weighted concepts	concept-based single vector VSM	Integrated in ranking	TFIDF	Cosine score

Table 2.2: Ontology-based IR models based in some adaptation of the VSM model

We could say that the work of [Fang et al., 2005] is a first try to include a semantic distance measure in an ontology-based IR model, although it be a coarse approximation, because the theory about ontology-based semantic distances described in section 2.2.3, offers a well founded and precise solution to this problem. Precisely, the Intrinsic Ontological Spaces model builds on previous results on this theory for providing a unified representation that integrates the intrinsic structures of the ontology in the model, providing many potential benefits to the common drawbacks of the family of ontology-based IR models revised throughout this section.

In [Mustafa et al., 2008], the authors propose a semantic IR model based in the use of RDF triplets and a thematic similarity function. The thematic similarity function associates concepts according to its membership in a common semantic field or theme. The user queries are encoded as RDF triplets, which are expanded to include synonyms and other semantically related concepts. The query expansion with related concepts uses a neighborhood notion based in a measure of semantic distance among concepts on the ontology. To establish the semantic similarity among the queries and the documents, the system uses the RDF triplets in the query and the RDF annotations associated to the documents. The documents with RDF triplets matching the terms in the expanded query are extracted from the collection, and are ranked according to their saliency. To select the documents that match the query terms, the authors use a set of semantic distance functions on the ontology to compute the closeness among the concepts in the query and the concepts annotated by the document, in other words, the retrieval of documents is driven by a ontology-based semantic distance function instead of a formal Boolean SPARQL query. To rank the retrieved documents, the documents are represented into a vector space of RDF concepts using a TFIDF weighting scheme, then, the documents are ranked using a combination of the cosine function and the same semantic distance function that was previously used. The semantic distance function used in the IR model is a novel edge-counting measure proposed in the same work, which includes a exponentially decreasing factor according to the depth of the nodes. The methodology can be summarized in four steps: (1) query expansion of the RDF triplets, (2) retrieval of related documents based in a novel edge-counting semantic distance, (3) mapping of the documents to a concept-based vector space using a TFIDF weighting, and (4) document ranking using the standard cosine function. The main drawback of the model of Mustafa et al. is that it retains the same geometric inconsistencies that previous ontology-based IR models, despite its smart integration of the semantic distances in the retrieval process. Although the model retrieves the documents using a ontology-based semantic distance, notion that we share as support in our model, in [Mustafa et al., 2008] the documents are ranked in a concept-based vector space where the semantic metric is missing. A second drawback of the model is the use of a edge-counting distance, which have been refuted by the research community, such as we discuss in section 2.2.3. Today, the most broadly accepted semantic distance is the Jiang-Conrath and its intrinsic variants. By last, other drawback of the IR model in [Mustafa et al., 2008] is that it does not consider instances of concepts, or named entities, in its representation, in contrast with the Intrinsic Ontological Spaces model proposed in section 2.4.

In [Dragoni et al., 2010], the authors propose a concept-based vector space model which uses WordNet leaf concepts as base vectors for the representation of documents

and queries. The proposed IR model is an adapted concept-based VSM model with an adapted TDIDF weighting method, and the standard cosine function as method for the ranking of saliency documents. The paper does not give details about the process to convert terms in WN concepts. Because the model does not include abstract concepts in its vocabulary, all the explicit references in the texts to abstract concepts not included in the vocabulary are discarded by the system. Also, the model does not include named entities recognizer (NER). Like the other concept-based adapted VSM models already described, the model of Dragoni et al. falls in the same modelling inconsistencies reported in section 2.1.1.

In [Egozi et al., 2011], the authors introduce a novel conceptual IR model based in the extension of a keywords-based VSM model with concepts defined in an ontological KB. Both, documents and queries are represented by a vector of weighted terms enriched with weighted concepts obtained through the use of an automatic annotation method, which extracts the underlying concepts within both text sources. The automatic semantic annotation method used is called *Explicit Semantic Analysis* (ESA) [Gabrilovich & Markovitch, 2006], and it is used to expand the standard terms-based VSM representation. The concepts used in the model are extracted from a hand-coded ontology. The authors use a feature selection method to choose the subset of concepts that best represents the corpus, and the selected concepts are used to expand the keywords-based VSM representation. The model proposed improves the results of previous methods when it is evaluated over some TREC corpus. By other hand, this model joins keywords and abstracts concepts in a same VSM model, thus, the authors follows the idea mentioned in [Castells et al., 2007] about the use of the ontology-based models as a complement to standard keywords-based models. The ESA model does not use a formal ontology to describe the structure relations of the concepts, although it could be easily extended to do it, such as is made by the authors in their proposal.

Moreover the common drawbacks of the family of ontology-based IR models, the main drawback of the model in [Egozi et al., 2011] is that it only includes references to abstract concepts (classes), not to entities (instances). From an abstract point of view, the model of Egozi et al. uses the same strategies that the models in [Castells et al., 2007], [Fang et al., 2005] and [Mustafa et al., 2008]. These strategies can be summarized as follows: (1) use a concept-based representation for documents and queries, (2) the use of ontologies, and (3) indexing and retrieval of documents based in a concept-based adaptation of the VSM model.

Unlike the model in [Castells et al., 2007], which builds two independent vector representations (keywords-based and concepts-based) that are combined later in the retrieval stage, the model in [Egozi et al., 2011] mixes concepts and meaningless terms in the same VSM representation, thus, the last one is an example of the *cardinality mismatch* problem described in the introduction of this chapter . Precisely, the core idea of the work in [Egozi et al., 2011] is to enrich the vocabulary based in keywords with concepts. The references to entities are captured by the meaningless keywords or terms, while the references to abstract concepts are captured through the ESA annotation method.

In [Cao & Ngo, 2012], the authors propose an extension of the keywords-based VSM model with ontological features associated to the named entities. The basic hypothesis is that the named entities are the more discriminative terms in most of

the user queries, therefore, the enrichment of the VSM model with information not explicitly represented in the documents should lead to improvements in the precision and recall measures. The main idea is to merge in a same vector representation the TFIDF weights derived from independent vocabularies with features from different nature. The model uses a multivector representation for each document, where each document is defined by a vector of TFIDF weights defined on multiples vocabularies associated to the different types of features, such as: keywords, the alias, the associated class to the named entity, and entity identifiers among others. By last, the model uses a barycentric combination[1] of the cosine function for each independent vector, such that the similarity between a document and a query is a weighted function of the individual similarities among pairs of independent feature vectors. The weight factors used to merge the independent similarity measures are left as free parameters to be tuned by each application. Other time, this adapted VSM model falls in the same modelling inconsistencies already reported.

In [Machhour & Kassou, 2013], the authors introduce a method to integrate the use of ontologies in VSM-based systems for text categorization (TC) already existent. The core idea of the method is to map the original term-based vectors, whose coordinates represent meaningless terms, to concept-based vectors whose coordinates represent concepts within ontology. The authors evaluate the proposed model with the known RCV1 corpus [Lewis et al., 2004], reporting only small improvements in performance, which they attribute to the strong pre-processing of these systems (stemming without disambiguation). Despite these discouraging results, the work studies a practical open problem with a clear application in TC.

2.2.2 Geometric representations for taxonomies

Our work is related in spirit with one distance-preserving ontology embedding proposed by Clarke in his thesis [Clarke, 2007], whose main ideas has been also published in [Clarke, 2009] and [Clarke, 2012]. Following some geometric ideas introduced by Widdows in [Widdows, 2004], Clarke proposes a distance-preserving embedding method for the concepts within a taxonomy, which is called *vector lattice completion*, whose main idea is to use the natural morphism between the taxonomies and the *vector lattices*.

Clarke's ideas are based in the very close relation between taxonomies and lattices, derived from the fact that many human-made taxonomies are join-semilattices, although in the more general case, we could also find examples of taxonomies with multiple inheritance, where a pair of concepts do not have a *supremum*.

The vector completion builds an order preserving homomorphism which maps each concept to a linear subspace in the vector lattice, with the property that the Jiang-Conrath distance among concepts [Jiang & Conrath, 1997] is preserved as the euclidean distance between vectors, when the taxonomy is a *tree*. The leaf concepts are mapped to base vectors of the space, while any non-leaf concept is mapped to the linear subspace spanned by its children concepts. We note that the ontology embedding of Clarke is an implicit application of the theory of categories [Pierce, 1991],

[1] A barycentric combination is a linear combination of variables defined by a set of weights, one per variable, whose sum is always one.

where his *completion* is a natural structure-preserving mapping among different, but intrinsically identical, algebraic structures.

Although the embedding proposed by Clarke represents a very important milestone in the search of a semantic distance-preserving representation for ontologies, and its application to the development of good ontology-based IR models, the work of Clarke has two important drawbacks in the context of an ontology-based IR model that differentiate it with the model proposed here: (1) the lack of the integration of individuals (instances of concepts) in the model, and (2) the lack of a method to represent information units composed by a collection of concepts or references to them, such as documents. Unlike our model, the Clarke's embedding does not consider populated ontologies, thus, the vector lattice completion only works for concepts, not for individuals (instances). Moreover, the model of Clarke cannot be used to represent information units defined by a collection of concepts, or references to concepts (instances), it means that we do not know how to use the vector lattice completion for representing and comparing documents. Precisely, Clarke surveys the compositionality vector-based representation problem in a recent work [Clarke, 2012].

The Jiang-Conrath distance and the valuation metrics. Clarke notes in [Clarke, 2007, pp.92] some relation between the Jiang-Conrath distance and one metric function on lattices, but he does not reach to unveil the key relation between the Jiang-Conrath measure and the definition of metrics on lattices. Clarke ask himself in [Clarke, 2007, pp.91] by the existence of other semantic distances as the Jiang-Conrath which allow to build other distance-preserving embedding using the vector completion approach. Now, we know that this property of the Jiang-Conrath is consequence of its relation with some sort of valuations on lattices, such as it has been investigated in [Orum & Joslyn, 2009]. The Jiang-Conrath is the metric associated to an antitone sort of valuation on a lattice.

Open questions about metrics and lattices. It is a known fact in lattice theory that any valuation function on a lattice induces a metric on it [Monjardet, 1981], by this reason, one possible research trend is to try to find a more general valuation definition which could subsumes any other ontology-based metrics. Other interesting question is to know if every metric on a lattice is a *valuation metric*, or there is other types of structures for them. In the hypothetical case that every metric on a lattice be a valuation metric, we could search an abstract valuation generalizing all the possible distance-preserving measures on a vector lattice completion.

2.2.3 Ontology-based distances

The necessity to compare semantic concepts has motivated the development of many semantic distances and similarity measures on ontologies. The distance and similarity functions are complementary functions with opposite meanings, in the sense that they produce an antitone, or inverse, ordering, it means that for a greater similarity decreases the distance and vice versa. By example, in the VSM model, most of models use the cosine function on the unit hypersphere (normalized vectors) which is exactly the inverse function of the geodesic distance among points on the features space, thus, despite of the cosine function or the geodesic distance produce an inverse ordering, they produce exactly the same ranking result for any input query.

Any similarity function can be converted in a distance function, and vice versa, thus, we focus here in the study of semantic distances on ontologies.

An ontology-based semantic distance d_O is a *metric* defined on the classes of any ontology, which verifies the definition 6. The ontology-based semantic distances can be categorized in three broad classes:

(1) Edge-counting based distances, such as [Rada et al., 1989], [Lee et al., 1993], [Wu & Palmer, 1994] and [Hirst & St-Onge, 1998].

(2) Vector-based distances, such as the measure in [Frakes & Baeza-Yates, 1992].

(3) IC-based distances. The family of distances based in the Information Content (IC) measure is the most broadly accepted one, whose main references are the works in [Resnik, 1995], [Jiang & Conrath, 1997] and [Lin, 1998]. The IC-based family is subdivided in two subgroups: (a) corpus-based measures which use corpus statistics to compute the occurrence probabilities and the IC values for each concept, and the (b) intrinsic methods which only use the information encoded in the structure of the ontology, in whose family we can cite the pioneering works of [Seco et al., 2004], [Zhou et al., 2008] and [Pirró & Seco, 2008].

The state of the art in semantic distances is defined by the IC-based measures [Meng et al., 2012]. The main research trend in the area is the development of intrinsic IC methods which use the intrinsic knowledge encoded in the ontology as means to avoid the computation of any corpus-based statistics. The research activity in intrinsic IC-based methods has increased very recently, such as is witnessed by a dozen of recently published papers.

According to some relevant benchmarks driven in the literature, we can conclude that the Jiang-Conrath semantic distance offers the best results for most of the applications, in special, whether its IC values are estimated by any intrinsic method. In [Budanitsky & Hirst, 2001], the authors carry-out some benchmarks to compare the IC-based measures of Resnik, Jiang-Conrath, Leacock-Chodorow, Lin and Hirst-St-Onge, concluding that the Jiang-Conrath distance offers the best results. In a later work [Budanitsky & Hirst, 2006], the same authors arrive to the same conclusion, and the work includes cites to other reports with similar conclusions about the JC distance. Finally, in [Sánchez et al., 2011] the authors carry-out a benchmark among IC-based measures comparing corpus-based methods and the more recent methods based in the computation of the IC values through intrinsic method. This last report concludes that all the measures work better using intrinsic IC computation, while the intrinsic JC distance get the second best global results for their tests.

Here, we only survey the most representative measures in the cited categories. For a broader revision of the literature, we refer to the reader to some recent surveys, some of them are focused in biomedicine, such as [Lord et al., 2003], [Lee et al., 2008], [Pesquita et al., 2009], [Hsieh et al., 2013], [Cross et al., 2013], and [Harispe et al., 2013], while others do not assume any specific domain, such is the case in [Saruladha et al., 2010], [Sánchez et al., 2012], [Xu & Shi, 2012], and [Gan et al., 2013]. The book by Deza and Deza also includes a short, but very

useful section about network-based semantic distances on ontologies as the Wordnet [Deza & Deza, 2009, §22.2].

2.2.3.1 Some history

The first ontology-based semantic distances to appear were the edge-counting based measures, whose main representative is the Rada's measure [Rada et al., 1989]. All these measures are characterized by the use of the shortest path length among concepts measured on the ontology graph. The key idea behind these methods is that the higher up you need climb to find a common ancestor to both concepts, the greater should be the distance between concepts, and vice versa.

In [Rada et al., 1989], the authors propose to use the shortest path length among concepts of an ontology as distance measurement among them, measure that they call *Distance*. Their work sets, up to our knowledge, the first known ontology-based semantic distance, and it also introduces the main hypothesis underlying all the subsequent ontology-based semantic distances: the *conceptual distance as metrics* hypothesis. This hypothesis states, following previous psychological studies, that the conceptual distance, or similarity, among concepts in a semantic network, is proportional to the path length that joins them. The shortest path length, also called *geodesic distance*, is a metric in the formal sense, by this reason, the authors in [Rada et al., 1989] prove that these measures are metrics on ontologies.

Hypothesis 1 (*Conceptual distance as metrics*) *The conceptual distance, or similarity, among concepts in a semantic network, is proportional to the path length that joins them [Rada et al., 1989].*

In table 2.3 we make a summary of the formulas used by some known measures to compute the semantic similarity, or distance, between a pair of concepts within an ontology, as well as the novel distance proposed in this work. The similarities appear as $sim\left(c_1, c_2\right)$, while the distance functions appear as $d\left(c_1, c_2\right)$. The function $de\left(c_i\right)$ returns the depth of any concept in the DAG of the ontology, it means the length from the concept to the root node. By other hand, function $L\left(c_1, c_2\right)$ denotes the shortest path length among two concepts.

The main drawback of the measures based in edge-counting is that they implicitly assume that every edge has the same relevance in the computation of the global path length, without to take into account its depth level or occurrence probability. This drawback can be called the *uniform weighting* premise. In [Resnik, 1995], the authors propose a new semantic distance based in a Information Content (IC) measure whose main motivation is to remove the *uniform weighting* premise of the edge-counting measures. The IC measure for every concept is only the negative logarithm[2] of the occurrence probability of the concept, such as is shown in (2.1). The probability function on the taxonomy defines a *probability space*, whose integral value on the taxonomy is 1. Resnik et al. define the similarity measure that is shown

[2]Throughout this work, the logarithm function included in the IC function is the binary logarithm $log_2(x)$, although it had been ommitted in the notation, such as is made in the rest of literature. In this case, the information content is measured in bits, like is usual for the entropy value.

Measure	Semantic similarity or distance measure
Rada	$d_{Rada}(c_1, c_2) = \min_{all\ paths} \{L(c_1, c_2)\}$
Wu-Palmer	$sim(c_1, c_2) = \frac{2de(LCA(c_1,c_2))}{de(c_1)+de(c_2)}$
Hirst-St-Onge	$d_{HS}(c_1, c_2) = \frac{L(c_1,c_2)}{k}$
Leacock-Chodorow	$d_{LC}(c_1, c_2) = \frac{L(c_1,c_2)}{\max_{c_i \in C}\{de(c_i)\}}$
Resnik	$sim(c_1, c_2) = \max_{c_i \in sup(c_1,c_2)} \{IC(c_i)\}$
Jiang-Conrath	$d_{JC}(c_1, c_2) = IC(c_1) + IC(c_2) - 2IC(LCA(c_1, c_2))$
Lin	$sim(c_1, c_2) = \frac{2ln(p(LCA(c_1,c_2)))}{ln(p(c_1))+ln(p(c_2))}$
weighted Jiang-Conrath (this work)	$d_{wJC}(c_1, c_2) = \min_{x \in Paths(c_1,c_2)} \left\{ \sum_{e_{ij} \in x} w(e_{ij}) \right\}$ $w(e_{ij}) = IC(P(v_i \mid v_j))$

Table 2.3: Examples of semantic distance and similarities

in table 2.3, which is equivalent to assign a weight with the value of the probability difference between the adjacent concepts of each edge.

Browsing the table 2.3, you can appreciate that other measures in the edge-counting family also are based in some combination of the shortest path value as was introduced by Rada et al, and all of them share the same drawbacks associated to the edge-counting family. It is the case for the works in [Lee et al., 1993], [Wu & Palmer, 1994], [Leacock & Chodorow, 1998] and [Hirst & St-Onge, 1998].

The key idea behind the IC-based distances is as follows. The probability function $p : C \rightarrow [0, 1] \subset \mathbb{R}$ is growing monotone while the ontology is bottom-up, thus, while we climb on the ontology, the observation probability of any abstract concept increase. As higher is the occurrence probability of one concept, lower is its information content and vice versa.

$$IC(c_i) = -log(p(c_i)) \tag{2.1}$$

In [Jiang & Conrath, 1997], the authors propose a set of IC-based semantic distances encoding a set of semantics notions that fill some gaps in [Resnik, 1995]. Jiang and Conrath follow the IC approach of Resnik, but they note that previous measures not consider some important semantic notions encoded by an ontology, which affects the semantic similarity appreciated by the human beings. They consider the following issues: the number of descendants, the global depth of the concepts, the type of semantic relation (hyper/hypo/meronymy), and the strength degree of a link between a parent concept and its children concepts. From the different measures proposed in [Jiang & Conrath, 1997], the wider accepted one is the Jiang-Conrath distance shown in table 2.3.

In [Lin, 1998], the author refutes the vector-based distances, such as the proposed in [Frakes & Baeza-Yates, 1992], by the necessity to use vectors, moreover, Lin also notes that the edge counting methods only works on taxonomies, not admitting

other ways of knowledge representation, such as first order logic. Lin propose a novel definition of semantic similarity based in a probabilistic model and the IC value.

2.2.3.2 Some facts about the Jiang-Conrath distance

The semantic distance proposed in [Jiang & Conrath, 1997] has two drawbacks that are solved by the novel ontology-based semantic distance proposed here, these drawbacks are as follows: (1) the Jiang-Conrath distance only is a metric in a strict sense when the ontology is tree-like, therefore the Jiang-Conrath does not satisfy the metric axioms on ontologies with lattice or general poset structure [Orum & Joslyn, 2009]; (2) the Jiang-Conrath distance is only uniquely defined for ontologies with lattice structure, not for those with a general poset structure, and (3) it is only defined on taxonomies of concepts, not weighted concepts (classes) or instances of concepts (individuals).

The standard formula of the Jiang-Conrath distance on taxonomies is given by the expression (2.2), where the term $LCA\left(c_1, c_2\right)$ means the *lowest common ancestor* node between the concepts c_1 and c_2. It can be written as $c_1 \vee c_2$ when the taxonomy is a join semilattice, because in this case every pair of concepts holds a supremum element.

$$d_{JC}\left(c_1, c_2\right) = IC\left(c_1\right) + IC\left(c_2\right) - 2IC\left(LCA\left(c_1, c_2\right)\right) \tag{2.2}$$

The formula (2.2) is uniquely defined for lattices, because in this case, we find that any pair of concepts shares a unique common ancestor, named supremum, and the third term is well defined. By contrast, for general taxonomies that not fulfill the lattice axioms, we find pairs of concepts with more than one lowest common ancestor, thus, the expression (2.2) admits more than a single value.

For the analysis in the next lines, we classify the taxonomies in three classes according to its structure as follows: (1) tree-like taxonomies (see fig. 2.2), (2) upper semilattices taxonomies (see fig. 2.3), and (3) general posets (DAGS) taxonomies (see fig. 2.4). Starting from the observations above, we summarize some of the main proven facts about the Jiang-Conrath distance as follows:

- *Up to date, the Jiang-Conrath distance has proved to offer likely the best results for a semantic similarity/distance measure.* This conclusion rises from many benchmarks carried-out in the literature, among we can cite the works in [Budanitsky & Hirst, 2006] and [Sánchez et al., 2012]. Today, the state of the art is based in intrinsic IC-based measures, in special, some intrinsic variants of the Jiang-Conrath measure, such as is reported in [Sánchez et al., 2012].

- *The Jiang-Conrath distance is a type of valuation metric on lattices.* A first glance to the formula in (2.3), found in [Deza & Deza, 2009, §22], induced us to think that the Jiang-Conrath distance is some sort of metric on a lattice, and the IC-values are some sort of valuation on it. We found that our first intuition was right. While we thought about it, we discovered that this way has been already walked by Orum and Joslyn [Orum & Joslyn, 2009]. The information content function defines an antitone valuation when the underlying taxonomy

is an upper-semilattice, and Orum and Joslyn prove that the IC function is a type of valuation on join-semilattices [3]. Precisely, the same authors explore the connection among semantic distances on ontologies and metric on lattices, proving useful theorems in the context of our work. For a survey about valuation metrics on lattices we refer to the reader to [Monjardet, 1981] and [Deza & Deza, 2009, §22].

$$v(x \vee y) - v(x \wedge y) = v(x) + v(y) - 2v(x \wedge y) \tag{2.3}$$

- *The Jiang-Conrath distance is uniquely defined only for taxonomies that verify the upper-semilattice structure, such as the tree-like ones.* Such as we explained above, this property is a consequence of the definition of the term $IC(LCA(c_1, c_2))$ as a function of the lowest common ancestor.

- *The Jiang-Conrath distance only is a metric on tree-like ontologies, not on semilattices or general posets.* The Jiang-Conrath is only uniquely defined on semilattices where every pair of nodes has a *supremum*, or unique Lowest Common Ancestor (LCA), however, in [Orum & Joslyn, 2009], Orum and Joslyn have proven that this condition is not enough to verify the axioms for a metric, because for some lattices or general rooted-posets (taxonomies) can happen that the triangle inequality not be satisfied. This theoretical result contradicts the claim made by Jiang and Conrath in their original paper [Jiang & Conrath, 1997], where they claim that their distance is a metric on any sort of taxonomy, without to include any exhaustive formal proof with regard it. This contradiction in absolute detracts anything to their work, and the deep impact that it has left in subsequent research.

- *The Jiang-Conrath distance is not uniquely defined on general taxonomies.* For the case of general posets, the Jiang-Conrath distance not only is not a metric, not even is well defined. The reason is that in this general case, it is possible the existence of pairs of concepts with more than one LCA concept. In a practical application, we can always select the first LCA concept found in a LCA search, but we conjecture that it can introduce discontinuities of distance function near of these elements, such as the discontinuity problems reported by Rada et al. in [Rada et al., 1989] as consequence of the constraint imposed in their distance function among sets of concepts.

- *The theoretical limitations of the Jiang-Conrath prevent to get a well founded metric space on general taxonomies.* One possible solution for the non uniqueness condition would be to compute all the LCA values for each pair of concepts [Baumgart et al., 2006], then, we could select the ancestral path with the minimum distance value as the Jiang-Conrath distance. This idea allows to define uniquely the Jiang-Conrath distance on any taxonomy, nevertheless, it is not enough to verify the metric axioms, because, such as was

[3] About this fact, we would like to clarify that there are some minor formal differences among the definitions given to the concepts "valuation" and "valuation metric", in [Monjardet, 1981], [Orum & Joslyn, 2009] and [Deza & Deza, 2009]. Thus, we think that would be needed a review of the definitions and results in [Orum & Joslyn, 2009] in light of the gaps in the literature.

proven in [Orum & Joslyn, 2009], it is not even possible in the simpler case of semilattices, where the uniqueness condition is already guaranteed.

- *The JC distance between one concept and its parent is equal to the difference of their information content values.* It means that any tree-like taxonomy endowed with the Jiang-Conrath distance can be interpreted as a weighted-graph where each edge is weighted by the IC difference between its adjacent concepts.

- *The JC distance between one concept and its parent is proportional to their join probability.* This fact is proven in [Jiang & Conrath, 1997], and it can be easily deduced, such as is shown in figure 2.5. Precisely, we use this fact to generalize the JC distance.

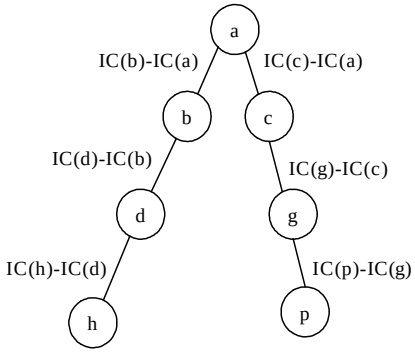


Figure 2.2: A taxonomy with tree structure

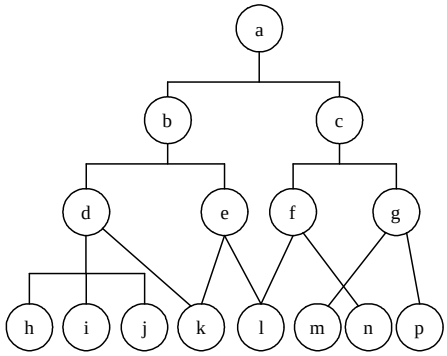


Figure 2.3: A taxonomy with upper semi-lattice structure

First, being as $c_2 \Rightarrow c_1$, and recalling the definition of joint probability, we get the expression in (2.4).

$$P\left(c_2|c_1\right) = \frac{P\left(c_2 \cap c_1\right)}{P\left(c_1\right)} = \frac{P\left(c_2\right)}{P\left(c_1\right)} \tag{2.4}$$

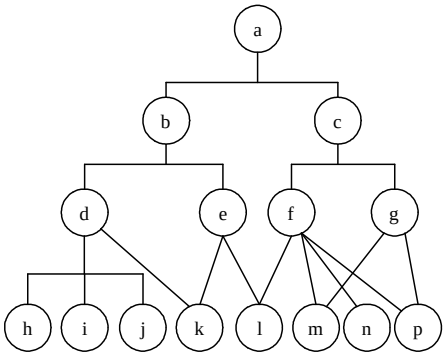


Figure 2.4: A taxonomy with general poset structure

Now, if we develop the JC distance among a parent concept c_1 and any child concept c_2, and we recall that $LCA(c_1, c_2) = c_1$, we get the expression in (2.5).

$$\begin{aligned} d_{JC}(c_1, c_2) &= IC(c_1) + IC(c_2) - 2IC(LCA(c_1, c_2)) \\ &= IC(c_1) + IC(c_2) - 2IC(c_1) \\ &= IC(c_2) - IC(c_1) \\ &= -log(P(c_2)) + log(P(c1)) \\ &= -log\left(\frac{P(c_2)}{P(c1)}\right) \\ d_{JC}(c_1, c_2) &= IC(P(c_2|c_1)) \end{aligned} \tag{2.5}$$

Because $P(c_2|c_1)$ is always less or equal to 1, the edge distances $d_{JC}(c_1, c_2)$ in (2.5) always take positive values, thus, we can define a weighted-graph on the taxonomy using these values as edge weights.

A *lowest ancestral path* is any path within the DAG (taxonomy) among two nodes which includes one, and only one common ancestor. Precisely, the Jiang-Conrath distance is defined by the three corners of any lowest ancestral path: the two extreme concepts and their lowest common ancestor. In this way, the JC distance can also interpreted as a IC-weighted shortest ancestral path, because if we add all the weights along one ancestral path we always get the value defined by the Jiang-Conrath distance, such as is expressed by (2.6). The weighted distance interpretation was already introduced in [Jiang & Conrath, 1997], and it can be appreciated in figure 2.5.

$$\begin{aligned} d_{JC}(c_1, c_2) &= \sum_{ij \in ancestral(c_1, c_2)} \Delta_{c_i, c_j} \\ &= \sum_{ij \in ancestral(c_1, c_2)} [IC(c_i) - IC(c_j)] \\ &= IC(c_1) + IC(c_2) - 2IC(LCA(c_1, c_2)) \end{aligned} \tag{2.6}$$

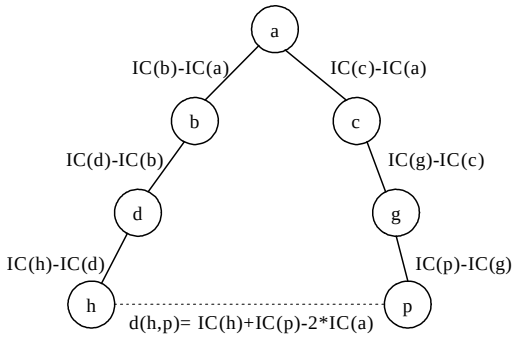


Figure 2.5: The JC distance interpreted as the length of the ancestral path defined by the weigthed edges.

The drawbacks of the Jiang-Conrath reported above motivate the development of the novel weighted Jiang-Conrath distance introduced in this work. Two interesting questions still are open: (1) why the JC distance is not a metric on a upper semilattice in spite of it is uniquely defined on it ?, and (2) is it possible to make some modifications to the basic JC formula to fulfill the metric axioms on any general taxonomy ?. In section 2.4.4, we answer the first question with an explanation, while we introduce a novel distance based in a generalization of the Jiang-Conrath distance as an answer to the second one.

2.2.3.3 Intrinsic IC-based distances

Today, it is well accepted by the research community that the IC-based semantic distance and similarities offer the best expected results in most of semantic evaluation tasks, however, the traditional IC-based family of methods has an important drawback from a practical point of view. The standard IC-based measures need to compute corpus-based statistics to evaluate the IC values for every concept within the ontology. The common method is to count every reference to a child concept as a reference to all its ancestors, and then using this frequency information to compute the occurrence probability for each concept on the ontology. The main problem with these corpus-based statistics is the difficulty to get well balanced corpus covering every concept in the ontology.

Motivated by the previous limitation, many authors have proposed novel methods, called intrinsic IC-based measures, whose main idea is to compute the IC values using only the information encoded in the same ontology, such as the density of the descendant nodes or its depth level respect to the root node. As pioneering works of this family, we can cite the works in [Seco et al., 2004], [Pirró & Seco, 2008] and [Zhou et al., 2008].

The number of intrinsic IC-based measures proposed has grown rapidly during the last five years, converting the area in the main research trend in semantic distance and similarity measures. Among the collection of novel proposals, we can cite the works in [Pirró & Euzenat, 2010], [KhounSiavash & Baraani-Dastjerdi, 2010], [Saruladha et al., 2011], [Sánchez et al., 2011] and [Sánchez & Batet, 2012], as well as the works in [Taieb et al., 2012], [Lingling & Junzhong, 2012], [Cross et al., 2013], [Harispe et al., 2013] and [Gupta & Gautam, 2014].

In spite of the huge research activity on the topic, the only available survey is the work in [Meng et al., 2012], although it is already out of date.

For the ontology-based IR model proposed in this work, we admit the use of any intrinsic IC-based method to compute the IC values, although our preferred method for this task is the proposed one in [Pirró & Seco, 2008].

2.2.4 Summary of the state of the art

All the ontology-based IR models revised fall in the category of concept-based adapted VSM models, with the exception of the model proposed in [Rada et al., 1989], which is based in the use of semantic metric spaces defined by one ontology-based semantic distances. The model of Rada et al. is very close in spirit and methodology to the Intrinsic Ontological Spaces model that we propose here, and we consider our model as a direct descendant of the pioneer work in [Rada et al., 1989]. Despite the great advances and results obtained by the family of ontology-based adapted VSM models, whose main representatives are the models of [Fang et al., 2005], [Castells et al., 2007] and [Mustafa et al., 2008], we think that the ontology-based IR models can be improved if the modelling inconsistencies shared by these models are solved, such as is proposed in this work.

Such as we saw in section 2.2.3, despite there are many semantic measures in the literature, it is broadly accepted that the Jiang-Conrath semantic distance offers the best results for most of the evaluated applications. The state of the art is to use the Jiang-Conrath measurement with some sort of intrinsic IC estimation, such as the methods proposed in [Seco et al., 2004], [Pirró & Seco, 2008] and [Zhou et al., 2008]. The current research trend about semantic distances is to develop novel intrinsic IC-based estimation methods and measurements. The Jiang-Conrath is very well founded thanks to its connection with the lattice theory, but it only defines a metric on tree-like ontologies, fact that is proven in [Orum & Joslyn, 2009].

In his thesis [Clarke, 2007], Clarke proposes a distance-preserving embedding method for the concepts within a taxonomy, which is called *vector lattice completion*, whose main idea is to use the natural morphism between the taxonomies and the *vector lattices*. Because most of taxonomies fulfill the join-semilattice axioms, the ideal completion builds an order-preserving homomorphism which maps each concept to a linear subspace in the vector lattice, with the property that the Jiang-Conrath distance among concept is preserved as the euclidean distance between vectors when the taxonomy is tree-like. The leaf concepts are mapped to the base vectors of the space, while any non-leaf concept is mapped to the linear subspace spanned by its children concepts. We note that the ontology embedding of Clarke is an implicit application of the theory of categories [Pierce, 1991], where his *ideal completion* is a natural structure-preserving mapping among different, but intrinsically identical, algebraic structures.

Despite the Clarke's model is not defined for individuals and it does not support the representation of set of concepts, his results are very important for our investigation, and his work has been a primary source of inspiration for us. In this regard, Clarke establishes an important theoretical result: he proves that a taxonomy can be embedded in a vector lattice, in such way that its topological structure (order)

and metric structure (semantic distance) be preserved.

2.2.5 Main differences with prior models

Next, we provide a summary of the differences between our ontology-based IR model, and the models reported in the literature.

- Unlike of the most of previous methods, our method represents the information units by sets of weighted-mentions to concepts (classes) or instances of concepts (individuals) within a metric space, instead of vectors whose coordinates represent weighted mentions on a set of mutually orthogonal vectors defined by the a set of concepts (classes) and/or instances of concepts (individuals).

- In our method, the mentions to concepts (ontological classes) are represented by sets with the following structure. Every image set in the representation space, associated to any class in the ontology, verifies the next property: the set subsumes all the subsets associated to the descendant classes (concept) and individuals (instance of concept) within the populated ontology, according to the metric space. By first time, a concept in the query is equivalent to the selection of a geometric subset of the representation space, that is, any logic query is converted in the selection of the geometric region containing all the concepts (classes9 and instances (individuals) subsumed by the concept cited in the query.

- Unlike other known methods, our method integrates in the same semantic representation space the mentions to concepts (classes) and instances of concepts (individuals) in a consistent way, through the preservation of the structures defined by the intrinsic geometry of the base ontology.

- Our method explicitly integrates and preserves the intrinsic geometry of the ontologies in the representation space, given by the next structure relations: (1) the order relation of the taxonomy, (2) its intrinsic semantic distance, and (3) the set inclusion for the individuals and subsumed concepts of the ontology.

- The weighted-mentions to concepts or instances of concepts are represented in a metric space based in a novel ontology-based semantic distance, in contrast with most of methods that uses a vector space model (VSM) and the cosine function as similarity measure. Our approach removes the implicit orthogonality condition associated to every VSM model, which is a source of semantic inconsistency in the previous representations.

- Unlike of previous methods, our method uses the Hausdorff distance as a metric on subsets of a metric space to compare and to rank information units (documents), instead of the cosine score. This feature also contributes to remove the implicit orthogonality condition of the VSM models. By other hand, the Hausdorff distance is a well defined metric on subsets of a metric space, which allows to remove the continuity problems reported in [Rada et al., 1989], and to build a semantic ranking function supported by a meaningful ontology-based distance, such as the novel distance introduced in our method.

- The proposed weighting method is defined as a statistical fingerprint, but it has a semantic meaning. The weight factor is a statistical and static derived from the frequency of every mention to a concept or instance within an information unit, equivalent to the standard TF weights used in all known IR models. However, the weight defines the ontology-based edge weight for each weighted-mention in the model, and it is a semantic weight defined by the IC-value of the mentioned ontological object. The weighting method proposed in this work combines, by first time, a statistical and static weight with an ontology-based semantic distance.

- The only known method that also uses a metric space for the representation of the information units is the model introduced in [Rada et al., 1989], but it presents some important differences respect to our method. First, the model of Rada et al. represents every document as a set of Boolean mentions to concepts, while our method includes a weighting method to represent the information units (documents) as a set of weighted-mentions to concepts and instances of concepts. Second, the model of Rada et al. uses the average ontology-based distance among concepts as a distance function among sets, while we use the Hausdorff distance, which is a strict metric among subsets and it allows the removal of some continuity problems reported by the authors in [Rada et al., 1989]. The ontology-based distance of Rada et al. does not include the distance among instances of concepts in its model, and it is based in the shortest path distance among concepts, while our method use the shortest weighted path distance among concepts with weights defined by a generalization of the Jiang-Conrath edge weights.

- Our method proposes a novel ontology-based semantic distance based in the shortest weighted path on the populated ontology, using the Jiang-Conrath edge weights, it means that every edge is weighted by the difference of IC-values in its extreme nodes. Our novel semantic distance is a generalization of the Jiang-Conrath distance, whose purpose is to remove the drawbacks described above. Unlike the standard Jiang-Conrath distance, our method is a well defined metric on any sort of ontology, while the first one is only a well defined metric on tree-like ontologies.

- Unlike of previous methods, our method defines a novel IR model where each one of its components is ontology-based, avoiding the loss of any semantic information derived from the base ontology of the indexing model. First, the representation space is defined by a metric space of weighted-mentions to concepts and instances, whose metric is ontology-based. Second, the weighting method, in spite to be a classical TF scheme, has a semantic contribution to the distance among items in the populated ontology, because the weigths define the joint probability for the weighted elements, whose IC-value is the length of edge joining any weighted-item to its parent concept/individual, thus, the weighting method is also ontology-based. Third, the ranking method is also ontology-based because it is based in the Hausdorff metric on subsets of the representing space, which derives directly from the ontology-based metric of the space. Fourth, the retrieval method is driven by the ranking method, this,

the retrieval operation is also ontology-based, Fifth, the information units are represented by a set of weighted-mentions to individuals and classes within ontology, therefore, the representation is directly defined on the underlying populated ontology space plus a metric derived from its structure. Sixth, the retrieval and ranking process is directly carried-out using the representation of information units, which avoids the necessity to interrogate the populated ontology through any formal query in SPARQL, or other equivalent language.

2.3 Preliminary concepts

The investigation carried-out in this work is inspired by a geometric point of view, and as consequence of this approach, we can appreciate the emergence of some geometric and algebraic structures, such as: metric spaces, differential manifolds, lattices, categories, topological spaces and so on. The readers familiar with this material can skip most of this section, nevertheless, for sake of completeness, we include it here to make easier the reading.

The topological spaces and the concepts of general topology are essential in every mathematical development on any kind of space, and they are usually included in most of undergraduate courses, by this reason, we refer the reader to the classic books of [Munkres, 2000] and [Arregui Fernández, 1988].

The motivation for the use of lattices and metric spaces in our discussion is their matching capacity to represent the intrinsic structure of the spaces studied and developed in this work. The main two types of spaces and models developed in this work are: (1) the taxonomic ontologies which integrate classes and individuals in a same metric space, and (2) a probabilistic model associated to each concept of a taxonomy, which allows us to use the semantic distances on ontologies based in the Information Content (IC) measure, such as the semantic distances introduced in [Resnik, 1995], [Jiang & Conrath, 1997] and [Lin, 1998].

2.3.1 Ontologies

The *Intrinsic Ontology Spaces* model only considers the *is-a* relations in an ontology, thus, we get a taxonomy of concepts, possibly with multiple inheritance, whose resulting model is a direct acyclic graph (DAG) with a root concept.

An ontology of taxonomic type, denoted by $\mathcal{O} = (C, E)$, is a direct acyclic graph (DAG) with a unique root node, where every element in the set C represents a concept and is named *node*, while the elements in the set E are named oriented edges (arcs). The edges in E encode the "is-A" semantic relations, known as hyperonymy and hyponymy. Our model is restricted to taxonomies, thus, we discard other semantic relations as the antonymy (opposite concepts) and meronymy (part-of), although the synonymy can be well integrated in the model if the ontology defines synsets as is made in Wordnet.

In spite of the limitations in this first version of our model, the structure of the taxonomies is enough to represent complex semantic relations, such as the multiple inheritance, thus its application scope is not excessively limited. The DAG structure of the ontologies define a *partial ordered set*, abbreviated by *poset.* The predecessors

of every node in the graph are named *ancestors,* and they represent concepts which subsume its *descendant concepts,* where these last ones represent more specific concepts of the first ones. We reserve the term "*parents*" for the first-degree ancestors of any node.

As example of an ontology, the figure 2.6 shows a small subset of semantic relations around the "armchair" concept in WordNet [Miller, 1995]. Wordnet is a very well known linguistic resource with ontological structure. Moreover the taxonomic type relations, Wordnet also includes other semantic relations among concepts, such as the synonymy and meronymy. As is shown in figure 2.6, the concept "seat" exhibits a large number of descendant concepts.

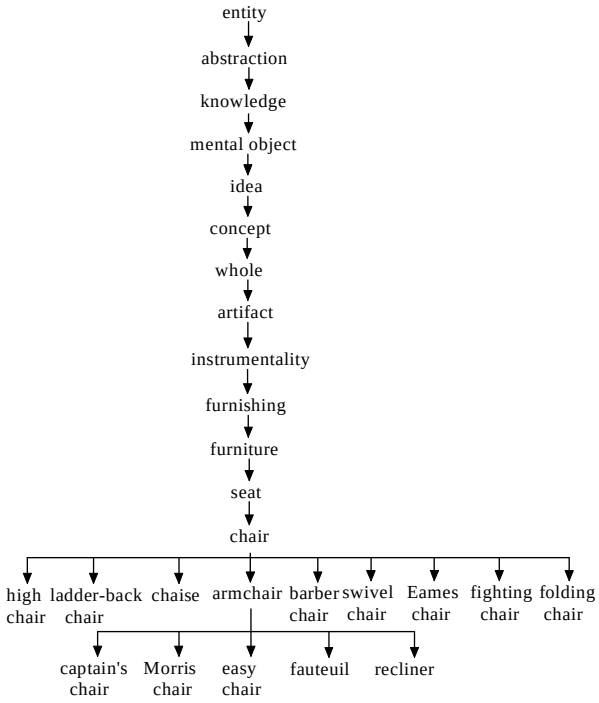


Figure 2.6: Partial sub-graph of WordNet around the "armchair" concept

As we said above, any ontology $\mathcal{O} = (C, E)$ is also a *poset* if we use the natural order relation on the DAG, which is denoted by $\leq_{\mathcal{O}}$, and its definition is given by (2.7). The acyclic condition of the DAG implies the antisymmetric property of the order relation $\leq_{\mathcal{O}}$, which is expressed as: if $(a, b) \in E \Rightarrow (b, a) \notin E$. For the definition of $\leq_{\mathcal{O}}$, we adopted the convention that the edges in the graph $\mathcal{O} = (C, E)$ are defined as $(parent, child)$.

$$b \leq_{\mathcal{O}} a \text{ when } (a, b) \in E \tag{2.7}$$

2.3.2 Lattices

Most of taxonomies, but not all, are tree-like, thus, they fulfill the axioms of a type of lattices named *upper-semilattices*, or *join-semilattices*. The upper-semilattices also allow to represent multiple inheritance taxonomies, but with a constrained structure: every two concepts must have a supremum. In algebra, the *lattices* rise

as a generalization of the partial ordered sets to other algebraic structures with some order intrinsic relation, in this sense, the lattices capture the common order relations among these structures.

The lattices allow to represent a great diversity of structures which share a common order structure. By example, the next structures are lattices: the family of linear subspaces of any vector space, partitions of subsets, the power set, Boole algebras, linear ordered sets, trees and directed graphs, integers ordered by the g.c.d and m.c.m. relations, and other cited in classical algebra books [Lidl & Pilz, 1998]. As a visual example, the figure 2.7 shows the lattice representation for the family of subsets of the finite set $\{1, 2, 3\}$.

A partial ordered set $(X, \leq)$, abbreviated as *poset*, is a non-empty set X where has been defined an *partial order* denoted by $\leq$, which is a binary relation that is *reflexive, antisymmetric* and *transitive*. The *order relation* is *partial* when not every pair of elements in the set is comparable. The *ordered sets* are well known structures whose properties can be consulted in any basic text about set theory, such as [Fernández Laguna, 2003].

Formally, the *algebraic lattices*, or simply *lattices*, are an algebraic structure that generalizes a type of ordered set named *ordered lattice*, whose definition is given below.

Definition 3 (Ordered lattice) *A poset $(L, \leq)$ is named ordered lattice if the supremum and infimum is defined for every pair of elements $x, y \in L$.*

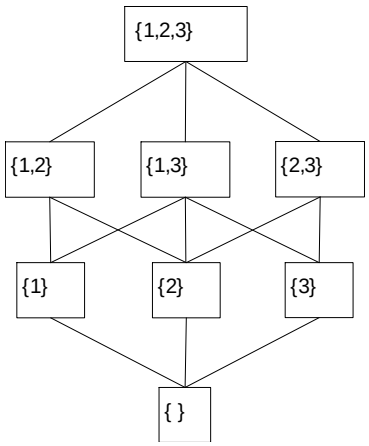


Figure 2.7: Lattice for the power set of {1,2,3}

The prior definition is generalized to other structures through the introduction of the abstract concept named *algebraic lattice*, or simply, a *lattice*. The generalization is made through the introduction of the generalized operators for the supremum and infimum, named as *join* ($\vee$) and *meet* ($\wedge$). In this way, we arrive to the abstract definition below [Lidl & Pilz, 1998]. For a broader study about lattices, we refer the reader to the classical book of Birkhoff [Birkhoff, 1967].

Definition 4 (Algebraic lattice) *An algebraic lattice, or simply, a lattice, is a tuple* $(L, \wedge, \vee)$ *with two binary operations denoted by* $\wedge$ *(meet) and* $\vee$ *(join), which verify the properties below for every set of elements* $x, y, z \in L$:

1. Commutativity: $x \wedge y = y \wedge x$, $x \vee y = y \vee x$
2. Associativity:$x \wedge (y \wedge z) = (x \wedge y) \wedge z, x \vee (y \vee z) = (x \vee y) \vee z$
3. Absorption: $x \wedge (x \vee y) = x$, $x \vee (x \wedge y) = x$
4. Idempotency: $x \wedge x = x$, $x \vee x = x$

When only the supremum (join), or infimum (meet) is defined for every pair of elements in a poset, we have a upper or lower semilattice structure, such as is defined below. The trees are examples of semilattices.

Definition 5 (Semilattice) *A upper-semilattice* $(L, \vee)$ *or lower-semilattice* $(L, \wedge)$*, is a tuple which verify the properties (1,2 and 4) above for every set of elements* $x, y, z \in L$.

Every algebraic lattice is isomorphic to any ordered lattice if the meet and join operators are defined, respectively, as the infimum and supremum on the associated poset, and turn, the supremum and infimum of the set can be recovered from the operators $\wedge$ and $\vee$. This important result is a basic theorem in lattice theory [Lidl & Pilz, 1998]. The lattice theory is very rich, and it has found application in many fields, thus, we refer to the reader to any classic book on the topic, such as [Birkhoff, 1963] or [Lidl & Pilz, 1998].

As a final remark about lattices, we would like to stress the generalization power of these structures. The lattices allow to build natural morphisms among different types of categories [Pierce, 1991], such as the examples cited above. Precisely, Widdows [Widdows, 2004] and Clarke [Clarke, 2007] base its geometric representations for taxonomies, in the fact that these semantic structures are upper semilattices and its elements can be represented as linear subspaces of a vector space. In summary, Widdows and Clarke build a mapping from the category of taxonomies to the category of vector spaces, thanks to the high level of abstraction in the relation encoded by the lattices.

2.3.3 Distances and metric spaces

The *metric spaces* are a type of *topological space* where a distance notion has been introduced, without to include any reference to a coordinate's system or any embedding space. The metric spaces are a generalization about the idea of distance among elements of the same set, which is omnipresent in geometry and many application fields, from there, its huge flexibility to model a great diversity of problems. Formally, a *metric space* is a set where a *function distance*, also called *metric,* has been defined for every pair of elements.

Definition 6 (Distance function or metric) *Given any non-empty set* X *and a binary function* $d : X \times X \rightarrow \mathbb{R}$*. The function* d *is named distance function, or metric, if for every set of elements* $\{a, b, c\} \in X$*, the function verifies the next axioms*

1. Positiveness: $d(a,b) \geq 0$
2. Coincidence or zero property: $d(a,b) = 0 \Leftrightarrow a = b$
3. Symmetry: $d(a,b) = d(b,a)$
4. Triangle inequality: $d(a,c) \leq d(a,b) + d(b,c)$

Definition 7 (Metric space) *A metric space is an ordered pair $\mathcal{X} = (X,d)$ where X is any non-empty set and $d : X \times X \to \mathbb{R}$ is a metric or distance function.*

A fast reading of the "Encyclopedia of Distances" [Deza & Deza, 2009], unveils us the huge richness and applicability of the distance functions and the metric spaces. Reading the book, we can appreciate that practically every algebraic structure, space, or structured set, admit some sort of *distance function*, allowing the definition of a metric space on them, and the use of all the properties derived from this powerful structure. By example, we can find distance functions defined over text strings, vector spaces, groups, lattices, or any sort of geometric objects as surfaces, curves or n-manifolds. The recent book by Deza and Deza is the main reference about the topic, being an invaluable resource in the study and modeling of any sort of metric space, thus, we encourage the reader to discover it.

2.3.4 Distances among sets

The distance functions allow to evaluate the dissimilarity among elements of a same set, but there are many situations where we need to measure the distance or dissimilarity among set of elements. It is the case in our work, where we are interested in the comparison among different information units which are defined by a set of semantic annotations.

The necessity to measure the distance among sets motivated the definition of the metric known as *Hausdorff distance*. The *Hausdorff distance* is a distance function according to the definition 6, but, instead of to be a binary function over elements of a set, it is a binary function among subsets of a metric space.

Given any metric space (U, d_U), we can define its associated *Hausdorff distance*, denoted by $d_H(x,y)$ in the following manner. First, we need to introduce the *point-set distance* concept. Next, the *Hausdorff distance* is defined as the supremum among all the point-set distances between elements of one set and the opposite one.

The distance between one element x of a set X and another set Y, called *point-set distance*, is defined as the minimum distance among x and all the elements in Y, using the metric on the universal set U, given by d_U.

Definition 8 (Point-set distance) *Be the pair (U, d_U) a metric space and $X \subset U$ any non-empty set in U. The distance from any element $a \in U$ to the set (subspace) X is denoted by $d_U(a,X)$ and defined by (2.8).*

$$d_U(a,X) = \inf_{x \in X} \{d_U(a,x)\} \tag{2.8}$$

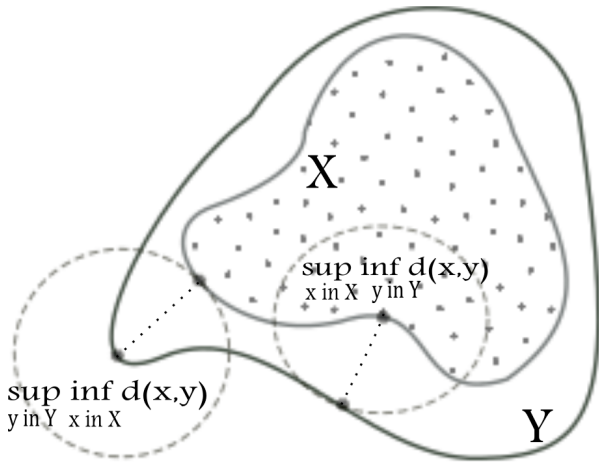


Figure 2.8: Hausdorff distance between subsets of a metric space

Finally, we arrive to the formal definition of the *Hausdorff distance* as is shown below. Note that the definition and evaluation of d_H depends on the original metric d_U associated to the metric space $\mathcal{U}$. The *Hausdorff distance* defined by d_H is a metric with the usual meaning given by the definition 6, but it is defined on the space of nonempty closed bounded subsets of a metric space [Henrikson, 1999]. In this way, the *Hausdorff distance* extends the metric d_U to enable the comparison among any pair of sets on a metric space. The Hausdorff distance is the maximum distance for each point-set distance value between the input sets. You can appreciate its geometric meaning in figure 2.8.

Definition 9 (Haussdorff distance) *Be the pair (U, d_U) a metric space and $X, Y \subset U$ two any non-empty closed bounded subsets in U. The Hausdorff distance between both sets is denoted by $d_H(X, Y)$ and defined by (2.9).*

$$d_H(X,Y) = max\left\{ \sup_{x \in X} \{d_U(x,Y)\}, \sup_{y \in Y} \{d_U(y,X)\} \right\} \tag{2.9}$$

2.3.5 Voronoi Diagrams

A Voronoi diagram is a partition of any metric space in subsets called *Voronoi cells*, which are defined as the set of closest elements to a distinguished element inside each cell. The distinguished elements are called *sites* or *centroids*. The Voronoi diagrams have many applications in fields where a space partition is useful for clustering or efficient retrieval, such as computational geometry, image processing or information retrieval. The Voronoi diagrams are formally defined below.

Definition 10 (Voronoi diagram) *Be (X, d_X) any metric space and $P = (p_k)_{k \in K}$ a tuple of elements of X called sites. A Voronoi diagram is the tuple of subsets $R = (R_k)$, called cells, where each cell R_k is defined as the subset of elements in X, whose distance to its associated p_k is not greater than their distance to other site in P, such as is expressed by (2.10).*

$$R_k = \{x \in X \mid d_X(x, p_k) \leq d_X(x, p_j) \ \ \forall j \neq i\} \tag{2.10}$$

2.4 Intrinsic Ontological Spaces

The *Intrinsic Ontological Spaces* are a sort of semantic representation spaces for populated ontologies, which are based in a metric space derived from an extension of the weighted Jiang-Conrath distance introduced in this work. These spaces represent semantically annotated information units, such as text documents, in a metric space composed by weighted-elements. The semantic annotations encode mentions to individuals or classes within a base ontology provided by the user. The base ontology is populated with new added data through an indexing and storing process.

The representation space is a metric space endowed with a hierarchical structure, which represents the classes (concepts) and individuals (instances) in a base ontology within the same space, while their intrinsic structure relations are preserved.

The intrinsic geometry of the ontology is defined by three structures: (1) the semantic distance (metric structure) among classes, (2) the taxonomic relations (graph/order structure), and (3) the set inclusion relations (set structure).

Our main goal is to design a semantic representation of the ontology that preserves these structures. The representation of the ontology is defined by a consistent metrization of the ontological space, integrating classes and individuals in a same representation, while their intrinsic structure relations are preserved. By consistent metrization, we refer to the definition of a metric that be used to compute any sort of distance among elements or subsets of the representation space, but it mimics the expected values of the original ontology-based semantic distance when its domain is constrained to the space of concepts on the base ontology.

The indexing method and IR model proposed in this work comprise the following components: (1) the definition of the semantic representation space as the universal set of weighted-mentions to individuals and classes within the populated base ontology, space that we call Intrinsic Ontology Spaces; (2) an embedding method to embed semantically annotated data, or information units, into the representation space; (3) an embedding method to embed semantically annotated queries into the semantic representation space; (4) a novel ontology-based semantic distance among concepts that we call weighted Jiang-Conrath distance, (5) a ontology-based distance among weighted elements (individuals and classes) of a populated base ontology, which defines the metric of the representacion space, and it is an extension of the weighted Jiang-Conrath ditance among concepts; (6) a ontology-based weighting method that combines statistical and semantic information to represent the semantic annotations associated to the indexed information units in the semantic representation space; (7) a novel ontology-based ranking method for the retrieval and sorting of the indexed units retrieved by the system; (8) a pre-processing step whose purpose is computing all the parameters and data structures to enable the indexing and searching operations of the search engine of the system; (9) an indexing and storing method to insert new data into the search and indexing system; and (9) a retrieval method to get a ranked collection of indexed units related to an input query;

In the remainder of the section, we describe the components of the model, as are enumerated above. We close the section with a summary of the objects and formulas defined by our model (see tables 2.4 and 2.5).

2.4.1 Notation and definitions

Throughout this section, we always use uppercase letters to denote sets, and lowercase letters for elements of a set. We start defining a taxonomy like a partially ordered set with a root node, and we continue adding structure to this poset.

Definition 1 (Taxonomy) *A taxonomy is a partial ordered set with a root node, denoted by the pair* $T = (C, \leq_C)$*, where* $C = \{C_i\}$ *is a non-empty finite set of concepts (classes) and* $\leq_C$ *denotes the order relation on the set* C*, and there is always a maximum element* $\rho \in C$*, called root, such that* $c_i \leq \rho, \forall c_i \in C$*. The set* C *is called set of concepts.*

If we endow the set of concepts C with any ontology-based metric $d_C : C \times C \rightarrow \mathbb{R}$, we get the metric space (C, d_C), structure that we call a *conceptual metric space.*

Definition 11 (Conceptual metric space) *A conceptual metric space* $\mathcal{C}$ *is the tuple* $(C, \leq_C, d_C)$ *where* $(C, \leq_C)$ *is a taxonomy and the binary function* $d_C : C \times C \rightarrow \mathbb{R}$ *is a metric on the set of concepts* C*.*

We can extend the taxonomy $T = (C, \leq_C)$ with a family of sets $I_C = \{I_{C_i}\}$ of instances (individuals) to concepts in C, to obtain the pair $(C \cup I_C, \leq_C)$, called a *populated ontology.*

Definition 12 (Populated ontology) *A populated ontology* O *is a tuple of type* $(C \cup I_C, \leq_C)$ *, where the pair* $(C, \leq_C)$ *is a taxonomy and* $I_C = \{I_{C_i}\}$ *is a family of sets of instances to the concepts in* C*. The family of sets of instances verifies the set inclusion relation with regard to its associated concepts, such that* $\forall C_i \in C \rightarrow I_{C_i} \subset C_i$*.*

By last, endowing the populated ontology with any ontology-based metric d_C, we get the tuple $\mathcal{O} = (C \cup I_C, \leq_C, d_{\mathcal{C}})$, structure that we call a *metric ontology.*

Definition 13 (Metric ontology) *A metric ontology* $\mathcal{O}$ *is a tuple* $(C \cup I_C, \leq_C, d_C)$*, where the pair* $(C \cup I_C, \leq_C)$ *is a populated ontology and the binary function* $d_C : C \times C \rightarrow \mathbb{R}$ *is a metric, such that the tuple* $(C, \leq_C, d_C)$ *is also a conceptual space.*

The goal of the IR model proposed in this work is to build a semantic representation for any corpus whose semantic annotation is based in a populated ontology $(C \cup I_C, \leq_C)$. We only consider the is-a relations within the ontology, discarding other sort of relations. The input to our IR model is a populated taxonomic ontology.

We note that the concepts C_i denote a set of elements with some common features, thus, every class associated to a concept defines a set by itself. The individuals of a class represent elements of the set denoted by its associated concept, and they satisfies the natural set inclusion relation, it means that every individuals set I_{C_i} satisfies the set inclusion relation, such that $\forall I_{C_i} \in I_C$ and $\forall C_i \in C$, it holds $I_{C_i} \subset C_i$.

We define the *ontological representation space* as the pair $\mathcal{X} = (X, d_X)$, with $X = C \cup I_C \times [0, 1] \subset \mathbb{R}$, where $d_X : X \times X \rightarrow \mathbb{R}$ is a metric on the space of *weighted-individuals.* The elements of the representation space are weighted references to instances of classes (individuals), or weighted references to concepts (classes).

Definition 14 (Ontological representation space) *An* ontological representation space $\mathcal{X}$ *is a pair* (X, d_X), *where* $X = C \cup I_C \times [0,1] \subset \mathbb{R}$ *is a space of* weighted-mentions to individuals and classes *within a populated ontology* $(C \cup I_C, \leq_C)$, *and* $d_X : X \times X \to \mathbb{R}$ *is a metric on it.*

In our model, any information unit is represented as a set of weighted references to classes or individuals. The inputs to the model are a collection of information units annotated with classes and individuals in the populated ontology $(C \cup I_C, \leq_C)$. The information units could be documents, or other information source that admits the same representation. The information units must have been semantically annotated with the frequency of typed entities (individuals) or classes within the base ontology. The weighting method used in the model consists in the unit normalization of the frequencies of every typed reference in the document (TF-weighting).

Definition 15 (Input information unit) *Any annotated information unit* δ_k *is a set of tuples* $\delta_k = \left\{ \left(\tau_j, f_j^k \right) \in C \cup I_C \times \mathbb{N} \mid j \in J(k) \right\}$ *where* τ_j *denotes the* j-th *reference to a class or typed individual within a populated ontology* $(C \cup I_C, \leq_C)$, f_j^k *the frequency of the concept or instance* τ_j *in the document* δ_k, *and* $J(k)$ *is a set of indexes of the individuals or classes cited in the information unit.*

Definition 16 (Space of frequency-based annotations) *The universal set* $D = C \cup I_C \times \mathbb{N}$ *of frequency-based mentions to individuals or classes within a populated ontology* $(C \cup I_C, \leq_C)$ *is called the space of frequency-based annotations on the base ontology* $(C, \leq_C)$.

Definition 17 (Space of annotated information units) *The tuple* $\mathcal{D} = \{\delta_k \mid \delta_k \subset D\}$ *defines the space of* semantically annotated information units. *Note that* $\mathcal{D}$ *is the space of subsets on* D, *it means its power set, such that* $\mathcal{D} = \mathcal{P}(D)$.

2.4.2 Design axioms

Once we have introduced the main elements of our IR model, we define the first principles, or axioms, that it should fulfill to bridge the gap identified as motivation of the present work. The intrinsic ontology embedding and Intrinsic Ontology Spaces are defined below, and they are also included in table 2.5.

Definition 18 (Intrinsic ontological embedding) *Given an ontology* $(C \cup I_C, \leq_C)$, *its associated* metric ontology $\mathcal{O} = (C \cup I_C, \leq_C, d_C)$, *a space of frequency-based annotations* $D = C \cup I_C \times \mathbb{N}$, *and a metric space* (X, d_X). *A functions pair* (φ_I, φ_C) $\varphi_I : D \to X$ *and* $\varphi_C : C \to X$ *is called an intrinsic ontological embedding, and the metric space* $\mathcal{X} = (X, d_X)$ *an Intrinsic Ontology Space, if the following axioms are satisfied. The function* d_H *is the Hausdorff distance among subsets on* (X, d_X).

1. Order (subsumption) invariance:

   (a) $C_1 \leq_C C_2 \Rightarrow \varphi_C(C_1) \subset \varphi_C(C_2), \forall (C_1, C_2) \in C \times C$

2. Metric invariance:

(a) $d_C(C_1, C_2) = d_H(\varphi_I(C_1), \varphi_I(C_2)) = d_X((C_1, 1), (C_2, 1)), \forall (C_1, C_2) \in C \times C$

3. Inclusion invariance:

(a) $\varphi_I(\tau) \subset \varphi_C(C_i), \forall \tau \in I_{C_i}$

(b) $\varphi_I(\tau) \subset \varphi_C(C_j), \forall \tau \in I_{C_i}, \forall C_j \mid C_i \leq_C C_j$

In this way, the Intrinsic Ontology Spaces are defined as an ontology representation space which verifies the structure-preserving axioms enumerated above. Note that in the proposed model, *every individual is considered as a child node of its parent concept*, with its own IC-value. By other hand, every whole class within the ontology is mapped as a set to the representation space through the function φ_C.

The definitions above are purely abstract, because they only defines the properties that the embedding functions and the Ontology Space must fulfill. The realization of the Intrinsic Ontological Spaces consists in the definition of these mathematical objects, task that is carried-out throughout this section.

Meaning of the axioms. The first axiom simply says that the mapping φ_C preserves the subset (order) relation among concepts in the representation space, while the taxonomy order is transformed in a space subset relation according to the metric d_X. The second axiom states the natural equivalence (morphism) among the input ontology-based metric d_C and the metric of representation space d_X, through the Hausdorff distance d_H on the space of subsets of X. Note that it simply means that the distance among concepts in the ontology is equal to the distance among its images, defined by the distance among its whole weighted-mentions $(C_1, 1)$ and $(C_2, 1)$. Finally, the third axiom states that the image of every individual (instance) of class in the representation space, must be included in the image set of its parent class and all the ancestor classes within the ontology. The last axiom (3.b) can be deduced from the axioms (1.a) and (3.a).

Why do we define these axioms as design principles ?. In prior paragraphs, we have defined our representation space in an abstract way. For sake of understanding, the reader can see the figure 2.9. The representation model that we propose is a mapping of a poset structure (taxonomy) into a hierarchical structure of metric subsets, which is topologically (order) and metrically equivalent (distance) to the original poset, plus the individuals annotated on it. The subsets in the taxonomy are defined by order relation, while the subsets in the intrinsic ontology space are defined by the metric of the space. We are converting an order relation (taxonomy) in a geometric relation (metric space).

2.4.3 Embedding for individuals, classes and info units

Embedding for individuals. Be $D = C \cup I_C \times \mathbb{N}$ the space of frequency-annotated information units, then the embedding function for individuals, denoted by φ_I, is given by (2.13). The function φ_I defines the embedding for isolated mentions to

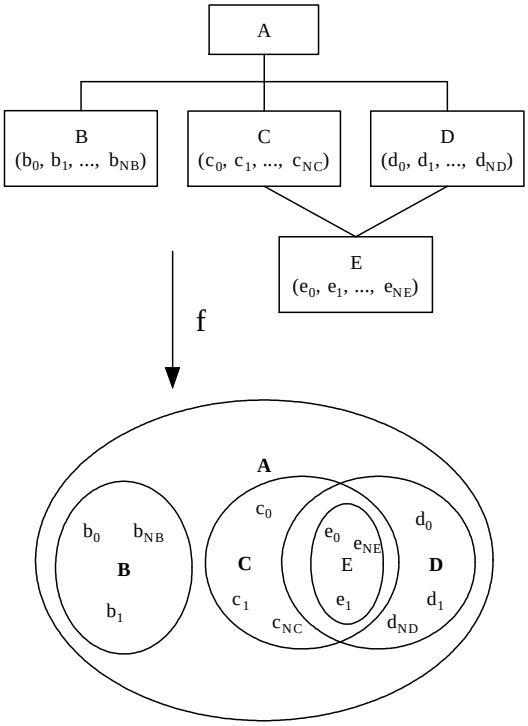


Figure 2.9: Intrinsic representation of any ontology

individuals or classes within the populated ontology in the metric space (X, d_X). For each class $C_i \in C$ exists a distinguished subset $(x_{C_i}, w) \subset X$, such that $(x_{C_i}, 1)$ is the image in X of a whole mention to the class C_i in any information unit. The element x_{C_i} is named the *site of the class* C_i.

$$\varphi_I \quad : \quad D \to X \tag{2.11}$$

$$\varphi_I\left(\tau_j, f_j^k\right) \quad = \quad \begin{cases} (\tau_j, 1), & \forall \tau_j \in I_C \\ (x_{C_i}, 1) & \forall \tau_j \in C \end{cases} \tag{2.12}$$

If one information unit contains only one mention to an individual or class within the populated ontology, the individuals get the static weight 1, without taking into account its frequency inside the indexed unit. Whether one information unit contains mentions to different individuals or classes, the function $\varphi_{\mathcal{D}}$ maps the information unit to a subset of the metric space (X, d_X), assigning a normalized TF weight based in the frequency of each item.

Embedding for documents. We recall that any information unit δ_k is a subset of the frequency-based mentions space D, such as is shown in the definition 15. The function $\varphi_{\mathcal{D}}$ maps a document δ_k, defined by a set of tuples of mentions to classes and individuals within the ontology, to a set of weighted mentions in the representation space. Such as you can appreciate in (2.16), the weights assigned to each mention are simply the normalized frequencies.

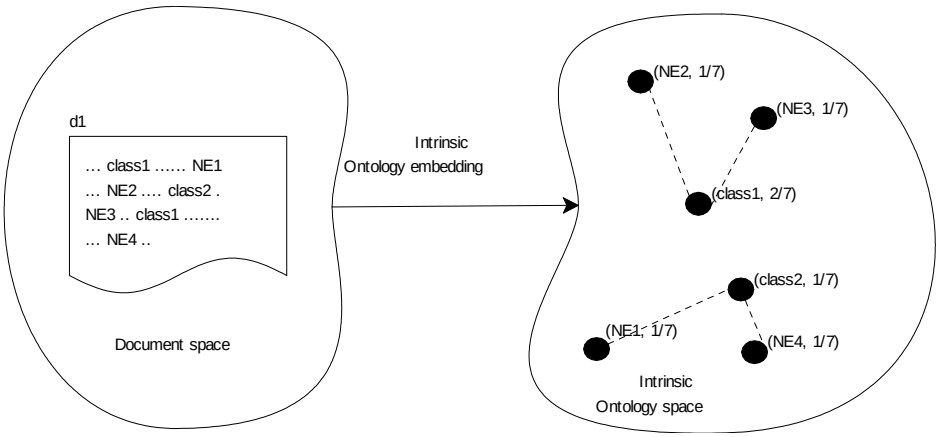


Figure 2.10: Document embedded in the ontological space.

$$\begin{aligned} \varphi_{\mathcal{D}} &: \mathcal{D} \to X && (2.13)\\ \varphi_{\mathcal{D}}(\delta_k) &\mapsto \left\{\left(\tau_j, w_j^k\right) \in X\right\} && (2.14)\\ \varphi_{\mathcal{D}}\left(\left(C_i, f_j^k\right)\right) &\mapsto \left(x_{C_i}, w_j^k\right) \in X, \forall C_i \in C && (2.15)\\ w_j^k &= \frac{f_j^k}{\sum\limits_{j \in J(k)} f_j^k} && (2.16) \end{aligned}$$

The function $\varphi_{\mathcal{D}}$ defines the embedding and static weighting method used to represent the information units (documents) in the representation space of the proposed IR model. The input to the function is a set of frequency-based weighted mentions to individuals and classes inside the information units to be indexed. The function $\varphi_{\mathcal{D}}$ computes a set of *static weights* for each mention through the normalization of the frequencies in the unit to be indexed, or represented, in the space (X, d_X) of the model. The function $\varphi_{\mathcal{D}}$ generates a set of tuples representing the weighted-mentions as indexing representation for the unit, such as is described in figure 2.10. Note that the ranking using these static weights combined with the semantic weights (normalized IC-values) to compute the distance between a query and the indexed documents.

We note a trivial fact, if any document $\delta = \{(\tau_j, f_j)\}$ contains a unique mention to one individual or class, then, $\varphi_{\mathcal{D}}(\delta) = \varphi_I(\delta) = (\tau_j, 1)$. The element $(\tau_j, 1)$ is the canonical whole mention to the individual or class τ_j.

Embedding for whole classes. In the case of queries, we consider that any mention to a class (concept) within the populated ontology is a reference to the whole class, it means that the query selects all the classes (concepts) and individuals (instances) subsumed by the mentioned class. Precisely, it is the meaning of the embedding function φ_C given in (2.17). The image of any class C_i is simply the image in the representation space X of every subsumed class or individuals within the ontology.

$$\begin{aligned} \varphi_C &: C \to X \\ \varphi_C(C_i) &= \{x \in X \mid \pi_O(x) \leq_C C_i\} && (2.17) \end{aligned}$$

The function π_O is a projection operator that takes a weighted-mention annotation $x \in X$, and returns the individual or class associated within a populated ontology $O = C \cup I_C$, such as is shown in (2.18).

$$\begin{aligned} \pi_O &: X \to C \cup I_C \\ \pi_O(\tau_j, w_j) &= \tau_j \end{aligned} \tag{2.18}$$

2.4.4 A novel ontology-based semantic distance

As we saw in section 2.2.3.2, the Jiang-Conrath is only a metric on tree-like ontologies. In the last paragraph of this section we introduced two questions that we answer below.

1. *Why the Jiang-Conrath distance is not metric on a upper semilattice in spite of it is uniquely defined on it ?.* First, we recall that the Jiang-Conrath distance is a weighted path metric associated to a selected *ancestral path* among two concepts on a tree-like ontology. Precisely, the reason because the Jiang-Conrath distance is not a metric on a lattice, or general poset, is that the JC distance is always selecting one lowest ancestral path, while it discard other alternative paths which could have a shortest weighted distance. In figure 2.11, we provide one example to illustrate this situation.

2. *Is it possible to modify in some way the Jiang-Conrath to get a metric on any sort of taxonomy ?* Yes, it is. Here, we introduce a novel generalization of the Jiang-Conrath distance that we call *weighted Jiang-Conrath distance*, which is defined as the shortest path on the weighted-graph associated to the taxonomy, using a generalization of the edge weights derived from the standard Jiang-Conrath distance on a tree-like taxonomy.

In figure 2.11, we show a taxonomy with lattice structure for which the distance d_{JC}, despite being uniquely defined, fails to be a metric. The weights on each edge correspond to the information content of the conditional probabilities, according to the standard Jiang-Conrath distance, such as was proven in the expression (2.5) above. Using the formula (2.2) with the binary logarithm for the IC value, we get the following distance values: $d_{JC}(c_5, c_4) = 8.016$, $d_{JC}(c_5, Object) = 2$ and $d_{JC}(c_4, Object) = 1.016$. We note that the function d_{JC} does not fulfill the triangle's inequality, because $d_{JC}(c_5, c_4)$ is greater than $d_{JC}(c_5, Object) + d_{JC}(c_4, Object)$.

The drawback of the distance function d_{JC} is the underlying selection of the lowest ancestral. The shortest path between the concepts 5 and 4 is defined by $[c_5, c_1, object, c_2, c_4]$, and this path induces an associated distance value $d_{wJC}(c5, c4) = 1+1+1+0.016 = 3.016$, in contrast with the prior value $d_{JC}(c_5, c_4) = 8.016$. Unlike the standard Jiang-Conrath distance, the novel generalized distance d_{wJC} always fulfills the metric axioms because it selects the shortest path on the weighted-graph. By other hand, the novel distance d_{wJC} matches the distance function d_{JC} when the taxonomy is tree-like.

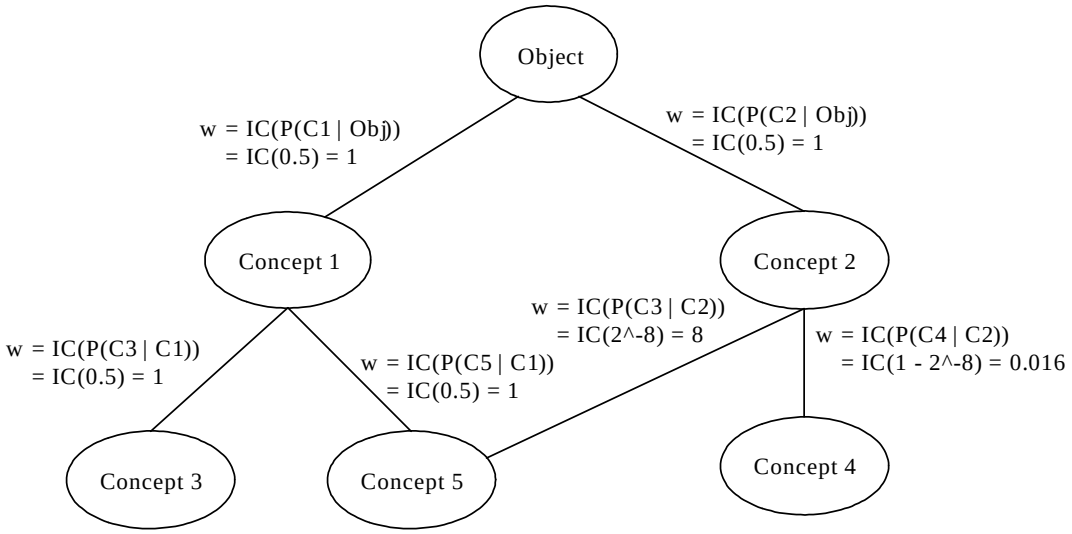


Figure 2.11: Taxonomy with lattice structure wherein the Jiang-Conrath fails to be a metric

2.4.4.1 Definition of the novel conceptual distance

The *weighted JC distance* is simply the shortest weighted path associated to the Jiang-Conrath edge weights, such as is shown by the expression (2.20). The shortest weighted-path distance is a metric on any sort of weighted-graph, and this measure allows to select any alternative path between two concept nodes to minimize its distance, not only the lowest ancestral paths. The novel distance transforms any taxonomy in a weighted-graph wherein the edge weights are the IC-values for the joint probability among the adjacent nodes to each edge. The distance is defined as the shortest path on the weighted-graph, and it can be computed using any known method to do it, such as any variant of the Djikstra's algorithm [Dijkstra, 1959] like [Ahuja et al., 1990], or many others available in the literature.

If the taxonomy is tree-like then the resulting distance function matches the standard Jiang-Conrath distance, otherwise, the weighted Jiang-Conrath distance always selects the shortest path according to the accumulated IC value throughout the edge path.

Definition 19 (Lowest common ancestor) *Given a partially ordered set $(C, \leq_C)$, the lowest common ancestor between two elements $a, b \in C$, denoted by $LCA\,(a, b)$, is defined by any common ancestor $x \in C$, such that does not exist other different element that also be an ancestor of the elements a, b. It is formally expressed by (2.19).*

$$LCA\,(a,b) = \{x \in C,\, a \leq x \text{ and } b \leq x \mid \nexists y \neq x \in C,\, a \leq y \text{ and } b \leq y\ \} \quad (2.19)$$

The LCA among concepts is unique for lattices, being called *supremum,* but it is not for general posets.

We define the undirected graph $G = (V, E)$ associated to the general poset $(C, \leq_C)$, where every vertex $v \in V$ represents an element in the set C, and one

edge $e_{ij} = (v_i, v_j) \in E$ exists if v_j is a lowest ancestor of v_i. A *path* between two different vertexes $a, b \in C$ is an ordered sequence of connected edges in the graph G, whose extremes match the vertexes a, b. We define the set of all the paths joining the elements $a, b \in C$ by $P(a, b)$.

The Jiang-Conrath distance, such as is defined in (2.2), has two drawbacks: (1) it is not uniquely defined for general posets, and (2) it is only a metric on tree-like taxonomies. To overcome these two limitations, we introduce a generalized version of the Jiang-Conrath distance that we call *weighted Jiang-Conrath distance.*

Definition 20 (weighted JC distance) *Given any base taxonomy $T = (C, \leq_C)$, we define a weighted graph $G = (V, E, w)$, with a positive real-valued function on each edge $w : E \to \mathbb{R}^+$. Every vertex $v \in V$ represents a element in the set C, and the edges are defined by $E = \{(v_i, v_j) \in C \times C \mid v_i \leq v_j \text{ and } v_j = LCA(v_i)\}$. The weighting function is defined by (2.21), and the weighted Jiang-Conrath distance is defined as the shortest weighted-path among two taxonomy nodes $a, b \in C$, such as is shown in (2.20).*

$$
\begin{aligned}
d_{wJC} &: C \times C \to \mathbb{R} \\
d_{wJC}(a, b) &= \min_{x \in P(a,b)} \left\{ \sum_{e_{ij} \in x} w(e_{ij}) \right\} && (2.20) \\
w &: E \to \mathbb{R}^+ \\
w(e_{ij}) &= IC(P(v_i|v_j)) = -log_2 P(v_i|v_j) && (2.21) \\
P(v_i|v_j) &\in [0, 1] \to w(e_{ij}) \geq 0, \forall e_{ij} \in E && (2.22)
\end{aligned}
$$

Despite the proposed generalization of the Jiang-Conrath distance looks obvious, it solves the problem, and up to our knowledge, it has not been introduced before. Our distance allows to define uniquely the value of Jiang-Conrath distance for general taxonomies, and getting a metric in strict sense, while it exactly matches its original definition for tree-like taxonomies, because in this case every pair of nodes only has a plausible path.

The definition of the weights $w(e_{ij})$ in (2.21) as a function of the joint probability $P(v_i|v_j)$ among one child concept and its parent guarantees that the values $w(e_{ij})$ are always positive. Note, that we have used $IC(P(v_i|v_j))$ instead of the difference of the IC values $IC_{vi} - IC_{vj}$, because this difference can be negative for some probability distributions on non tree-like taxonomies. We also note that the weights in (2.21) only match the IC difference, as is proven in (2.5), when the ontology is tree-like, because otherwise, the occurrence probability for nodes with more that one ancestor induces a IC-value that does not match this IC difference.

The evaluation of the function d_{wJC} requires the implementation of any shortest path algorithm, which could be time-consuming for large taxonomies. Due the probabilities of the concepts are known a priori (intrinsic IC), and the individuals are inserted in the taxonomy as children of its parent concepts, it is possible to compute a priori all pairwise distances in a pre-processing step, such as is shown in figure 2.13.

2.4.4.2 Extension to the whole representation space

Our model integrates individuals and classes in the same representation space. Moreover, the individuals are represented in the space as weighted-mentions instead of whole mentions (boolean) to entities. All the known ontology-based distances, such as $d_{weightedJC}$, are only defined for the concepts within the taxonomy, however, we need to extend this distance to the space of weighted-mentions. First, we must to define how the individuals are represented in the model, and how is extended the distance to include them in its domain. Second, we need to extend the distance d_{wJC} to the space of weighted-mentions, while we guarantee that the distance value among concepts (classes) mimics the function $d_{wJC}|_C$ constrained to the space of concepts C.

The *Intrinsic Ontology Spaces* defines a metric space of weighted-individuals as representation space for the semantically annotated data, and it considers four types of weighted-mentions: (1) mentions to whole classes (set of subsumed concepts and individuals), (2) mentions to whole individuals (whole or boolean instances), (3) weighted-mentions to individuals (partial instances), and (4) weighted-mentions to classes (partial concepts).

Our aim in this section is to extend the novel distance d_{wJC} in (2.20) to the elements of our representation space. The key idea to achieve the proposed goal is to build a consistent extension of the function d_{wJC}, in such way, that the resulting function d_X can measure the distance among the four types of elements cited above, while it matches the function d_{wJC} constrained to the space of concepts C. Because the distance d_{wJC} is strongly coupled to the structure of the taxonomy, the natural solution, and almost obvious, is to represent all sort of elements in the representation space ia an extension of the base taxonomy, such as is shown in figure 2.12.

The figure 2.12 offers a visual representation of the meaning of the metric d_X which defines the metric space (X, d_X) that constitutes the core of our IR model. The *instance 1* represents a whole instance of the concept 3, which is considered as a child node of its parent concepts. Any weighted-individuals $(inst, w_{inst})$ is considered as a child node of the whole individual, and the same holds for the weighted-mentions to classes (x_{c_i}, w_{c_i}). Our model admits that any individual belongs to more than one class, such as is shown in figure 2.12.

Extending the conceptual distance to individuals. In our model, every individual is considered as a child node of its parent concept, such as is shown in figure 2.12. The distance between two individuals is always the sum of the distance among its parent concepts plus the distances from each individual to its parent. Therefore, the distances among concepts (classes) do not change once the base ontology is defined, allowing the use of the precomputed concept-concept distances, such as is shown in figure 2.13. In this way, the distance d_X among weighted-mentions in X is defined by (2.24) and (2.26).

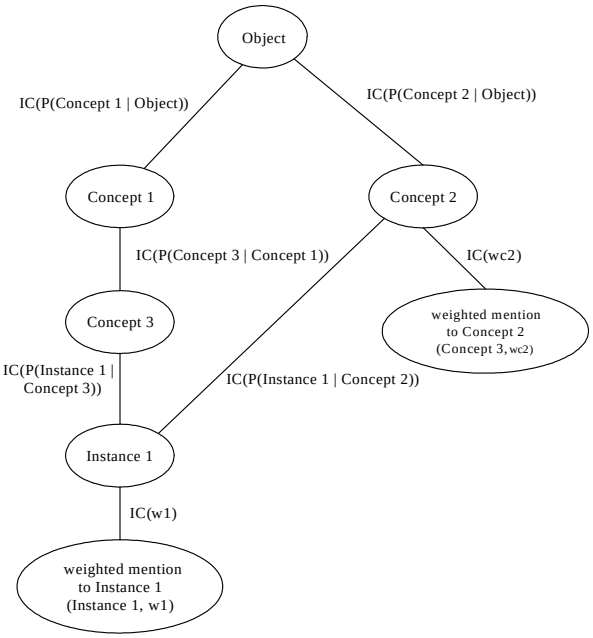


Figure 2.12: Unified representation of weighted and whole mentions to individuals and classes in a same taxonomy.

$$LA \quad : \quad C \cup I_C \rightarrow \mathcal{P}(C) \tag{2.23}$$

$$LA(x) \quad = \quad \begin{cases} x, & \text{if } x \in C \\ \{c_i \in C \mid x \leq c_i, \nexists c_j \neq c_i \text{ y } x \leq c_j \leq c_i\} \end{cases} \tag{2.24}$$

$$d_X \quad : \quad X \times X \rightarrow \mathbb{R} \tag{2.25}$$

$$d_X((x, w_x), (y, w_y)) \quad = \quad \begin{cases} \min\limits_{LA(x) \times LA(y)} \left\{ \begin{array}{l} \underbrace{-log(w_x \cdot P(x|LA(x)))}_{(1)} \\ \underbrace{-log(w_y \cdot P(y|LA(y)))}_{(2)} \\ \underbrace{+d_{wJC}(LA(x), LA(y))}_{(3)} \end{array} \right\} & , \text{ if } x \neq y \\ \left| log\left(\frac{w_x}{w_y}\right) \right| & , \ x = y \end{cases} \tag{2.26}$$

The terms (1) and (2) in the formula (2.26) are defining the edge weights and semantic distance among every whole instance (individual) and its parent concept (class), denoted by $LA(x)$, where LA means *lowest ancestor concept*. These terms also correspond to the information content of the conditional probabilities in (2.27), wherein $x \in I_{C_{LA(x)}}$ is an individual belonging to the class $LA(x) \in C$.

$$IC(x) - IC(LA(x)) = -log(P(x|LA(x))) \tag{2.27}$$

The weights $w_x, w_y \in [0, 1]$ in the terms (1) and (2) of the expression (2.26) correspond to the static frequency-based weights associated to the weighted-mentions (x, w_x) and (y, w_y), such as are given by (2.32).

The binary function d_X is also a shortest path metric on the extended ontology. It is inmmediate to prove that if $d_{wJC} : C \times C \to \mathbb{R}$ is a metric, then $d_X : X \times X \to \mathbb{R}$ is a metric on the representation space X.

Extending the conceptual distances to weighted-mentions to individuals or classes. In formula (2.26), the terms (1) and (3) are measuring the distances from the weighted-mentions to its whole instances nodes, denoted by $x \in I_{C_{LA(x)}}$ and $y \in I_{C_{LA(y)}}$. We are interpreting the weighted-mentions (x, w_x) and (y, w_y) as two virtual subsumed instances $x' \subset x$ and $y' \subset y$ of the parent concepts x and y, it means that we consider a partial (weighted) mention to an instance as a subsumed concept of the whole instance. This interpretation allows to integrated all sort of partial (weighted) or whole mentions to individuals and classes in a same representation space, while the metric of the space is a consistent distance function with regard to the base ontology-based distance d_{wJC}, and the different types of elements of the space, such as: the weighted-mentions (mention to a partial object), the whole instances of concepts and the parent concepts subsuming them. By last, the weights $w_x$and w_y assigned to the mentions to the instances x, y, can be interpreted as the conditional probabilities in (2.28).

$$\begin{aligned} w_x &= P\left(x'|x\right) \\ w_y &= P\left(y'|y\right) \end{aligned} \tag{2.28}$$

Precisely, because the edge weights in the weighted-graph associated to the Jiang-Conrath distance are the information content of the conditional probability, such as was shown in (2.5), the edge weights corresponding to the edges joining a weighted-individual (x, w_x) to its referenced whole individual $(x, 1)$ can be defined as the information content of this conditional probability, such as is shown in (2.29).

$$\begin{aligned} d_{wJC}\left(\left(x, w_x\right), \left(x, 1\right)\right) &= IC\left(P\left(\left(x, w_x\right) | \left(x, 1\right)\right)\right) \\ &= -log\left(w_x\right) \end{aligned} \tag{2.29}$$

2.4.5 Ontology-based weighting

The documents and queries are defined as sets of weighted mentions to individuals and classes, but with one difference, any mention to a class in a query is considered as a reference to all the classes and individuals subsumed by this class. The weighting scheme is simply the unit normalization of the frequencies associated to every semantic annotation of the documents, such as is explained below.

For each document, a semantic annotation method is used to automatically identifying the mention to typed individuals and concepts in the ontology. The semantic annotations are inserted as individuals in the ontology, jointly with one weight and cross-reference for the container document.

Given a document $\delta_k = \left\{\left(\tau_j, f_j^k\right)\right\}$, and assuming a i.i.p. model, we note that the probability $P\left(\delta_k\right)$ is given by the expression (2.30) ,where n_j^k is the number of occurrences of the individual or concept τ_j in the document δ_k.

$$P\left(\delta_k\right) = \prod_{j \in J(k)} P^{n_j^k}\left(\tau_j\right) \tag{2.30}$$

Doing some algebra from (2.30), we arrive to the expression (2.31) for the information content of the document δ_k. The meaning of (2.31) is that the information content of any document is equal to t he sum of the information contents of the semantic annotations, weighted by the frequency of each annotation.

$$\begin{aligned} IC\left(\delta_k\right) &= -logP\left(\delta_k\right) = -log\prod_{j\in J(k)} P^{n_j^k}\left(\tau_j\right) \\ IC\left(\delta_k\right) &= \sum_{j\in J(k)} n_j^k IC\left(\tau_j\right) \end{aligned} \tag{2.31}$$

A basic requirement of any IR model is to be able to compare documents with different lengths, but equivalent semantic fingerprints. Therefore, we normalize the value of $IC\left(\delta_k\right)$ to get the normalized information content, denoted by $\widehat{IC}\left(\delta_k\right)$ in (2.32).

$$\widehat{IC}\left(\delta_k\right) = \sum_{j\in J(k)} w_j^k IC\left(\tau_j\right), \qquad w_j^k = \frac{n_j^k}{\sum\limits_{j\in J(k)} n_j^k} \tag{2.32}$$

The expression (2.32) supports the definition of our TF weighting scheme for documents in (2.33). A document $\delta_k = \left\{\left(\tau_j, f_j^k\right)\right\}$ is embedded in the representation space as a set of weighted mentions to typed individuals or classes within the ontology, while in the case of the queries, all the weights for individuals or classes take the value 1, and the mentions to classes are mapped to full classes through the mapping φ_C.

$$\varphi\left(\delta_k\right) = \left\{\left(\tau_j, w_j^k\right) \mid j \in J\left(k\right)\right\} \tag{2.33}$$

The mapping $\varphi\left(\delta_k\right)$ for any document defines it as a *barycentric combination* of its semantic annotations. The weights w_j^k define a set of static values used to build the index form of ontology-based annotated data. These weights are defining an ontology-based semantic distance according to the metric of the space, given by the expression (2.26).

2.4.6 Ontology-based ranking

Given two indexed documents $\delta_k, \delta_m \in \mathcal{D}$, or information units, the distance in the input space $\mathcal{D}$ is denoted by $d_{\mathcal{D}}$ in (2.34), and it is simply the Hausdorff distance among subsets in the representation space (X, d_X), such as is shown in (2.35).

$$d_{\mathcal{D}} : \mathcal{D}\times\mathcal{D}\to\mathbb{R} \tag{2.34}$$

$$d_{\mathcal{D}}(\delta_k,\delta_m) = d_H\left(\varphi_{\mathcal{D}}\left(\delta_k\right), \varphi_{\mathcal{D}}\left(\delta_m\right)\right) \tag{2.35}$$

Given any query $q \in \mathcal{D}$, the ranking method is reduced to the computation and sorting of all pairwise distances $d_{\mathcal{D}}(q, \delta)$ among the query and every indexed unit δ. It means to compute the Hausdorff distances among the image of the query and the image of every indexed document.

The ranking method is a search and sorting process for the nearest neighbors in the representation space. For large collections, this process can be optimized using some sort of geometric partition on the representation space, such as the method proposed by Brin in [Brin, 1995], or other known methods. The optimization of the geometric search is a known problem in the literature, which is out of the scope of our research, however, because our model is a well defined metric space, there are many known methods that could be directly integrated in our model.

Open problem 1 (Ontology-based space partition) *Following our structure-preserving approach, we wonder if would be possible to develop some novel space partition and search structure based in the hierarchical structure of our representation space, approach that we could call as* ontology-based space partition*. Note that the weighted-mentions included in our space are already partitioned according to the ontology structure. Here, the question is if we can define some kind of partition for subsets which takes advantage of this property, instead of use a partition for general metric spaces, such as is proposed in [Brin, 1995].*

For the evaluation of $d_{\mathcal{D}}(q,\delta)$, we must aware that any mention to a whole class $C_i \in C$ in the input query q, has $\varphi_C(C_i)$ as its image in the representation space, such as is given by (2.17). The image of any whole class defines a subset (region) of the representation space, and we can use its definition to simplify the computation of $d_H(\varphi(q),\varphi(\delta))$ when it is necessary to compute the distance between any weighted-mention to class or individual in $\varphi_{\mathcal{D}}(\delta)$, and the image of any whole class $\varphi_C(C_i) \subset \varphi(q)$ in the query. In this case, the distance among any weighted-mention and a whole class C_i is given by d_{IC_i} in (2.36).

$$\begin{aligned} d_{IC_i} &: X \times \varphi_C(C_i) \subset X \to \mathbb{R} \\ d_{IC_i}\left(\left(\tau_j,\omega_j^k\right),\varphi_C(C_i)\right) &= \begin{cases} 0 & \text{, if } \tau_j \leq_C C_i \\ d_X\left(\left(\tau_j,\omega_j^k\right),(x_{c_i},1)\right) & \text{, otherwise} \end{cases} \end{aligned} \tag{2.36}$$

2.4.7 Pre-processing step

In figure 2.13, the reader can see a flowchart of the pre-processing step of the proposed IR model, whose main goal is computing all the static parameters required for the operation of the proposed IR model. These parameters are as follows: (1) the Information Content (IC) values for each ontology node, (2) the weights for the ontology edges, and (3) all pairwise semantic distances among the classes of the ontology. The all pairwise semantic distances correspond to the values of the function $d_{weightedJC}$ in (2.20).

The input data to build the IR model is a base ontology in any valid file format, such as an OWL file in XML format. The IC values could be obtained through corpus statistics, or using any of the intrinsic methods cited in section 2.2.3.3. Our favorite approach is to use an intrinsic method, such as [Pirró & Seco, 2008], because it only depends on the topology of the ontology, although we should never lose of sight that all these methods is trying to infer the joint probabilities among concepts through the logic building process of the taxonomy, therefore, there will be always different plausible and valid methods to do it, and we define our IR model as agnostic regard to them.

2.4.8 Indexing process

In figure 2.14, we introduce a flowchart for describing the indexing method for any novel information unit to store in the system ontology-based repository. First, any user provides an information unit, such as a text document (box 1). Second, the same automatic semantic annotator of the query process (box 2) is used to identify the mentions to individuals and classes which are present, or able to be represented, in the base ontology. The semantic annotation step produces a set of semantic annotations plus its frequency within input document. Third, the document is embedded (represented) in the Intrinsic Ontological Spaces as a set of normalized weighted-mentions to classes and individuals in the base ontology. Fourth, the indexing step is split in two steps: (1) the first step in box 5 has as main aim the storing of the index form, defined by a set of static semantic weights, into the repository for the indexed information units (box 7); (2) the second step (box 6.1) has as main goal the storing of the semantic annotations of the input document into the populated base ontology (box 8), then, the new annotations are extended to keep the back reference (inverse map) to the indexed units where it appears (box 6.2); and the IC-values for the registered individuals are updated (box 6.3).

2.4.9 Retrieval process

In figure 2.15, we introduce a flowchart for describing the method to compute a set of ranked information unit as answer to an input query provided by any user. First, the user provides an input query in text format, or other symbolic representation (box 1). Second, the system uses any automatic semantic annotator (box 2), out of the scope of this work, to convert the input query in a set of semantic annotations to individuals or classes within the populated base ontology of the system (box 3). Third, the ontology-based representation of the query is embedded in the Intrinsic Ontological space (box 4) as set of weighted-mentions to individuals or classes within the base ontology. Fourth, in the box 5 we can appreciate the process to retrieve and ranking the indexed information units (documents) respect to the input query.

The retrieval and ranking is based in the Hausdorff distance d_H among the query and the indexed units, such as is defined by the function $d_{\mathcal{D}}$ in (2.34). By other hand, the distance d_H is defined by (2.9), being derived from the metric d_X of the representation space (X, d_X) defined by the formula (2.26).

By last, the system sorts the indexed documents according to the semantic distance to the input query, and it returns a set of ranked information units.

2.4.10 Hierarchical Voronoi diagram

It is easy to prove that if the ontology is tree-like, then the *ontology embedding* defined in (2.13) is a *hierarchical Voronoi diagram* (HVD) on the classes of the

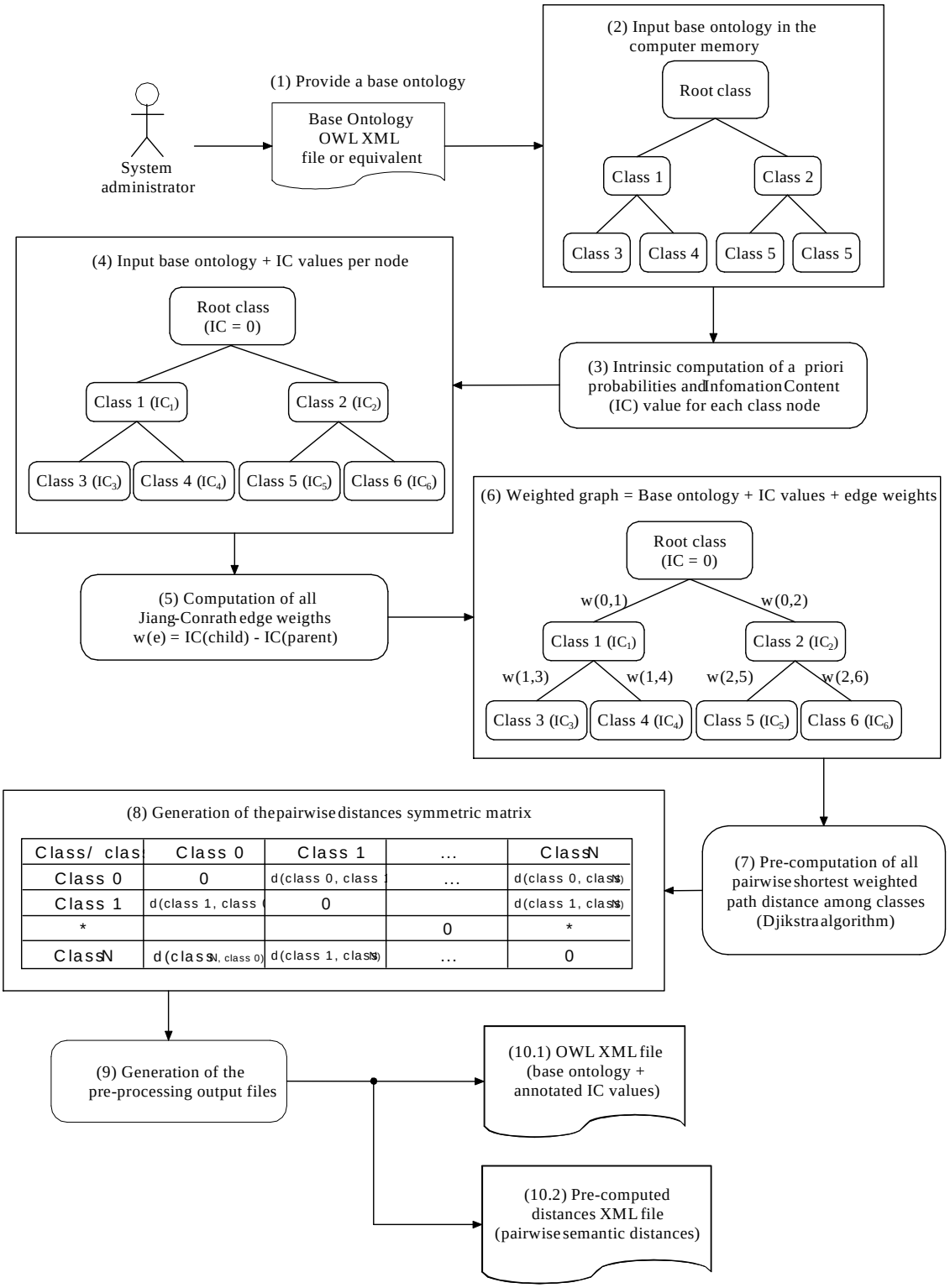


Figure 2.13: Pre-processing step for the computation of all pairwise distances among concepts, and the concept IC-values

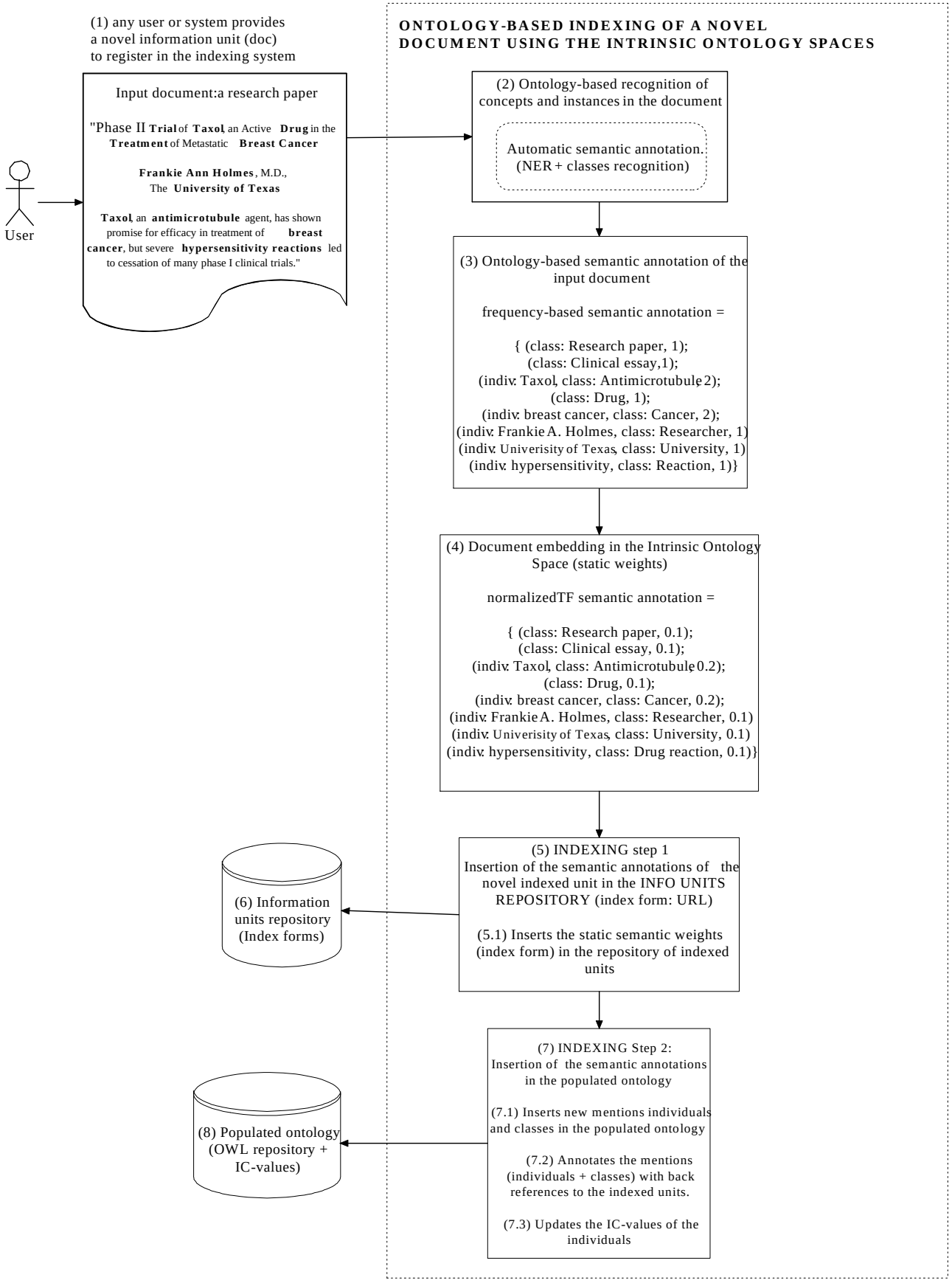


Figure 2.14: Indexing process for a novel information unit

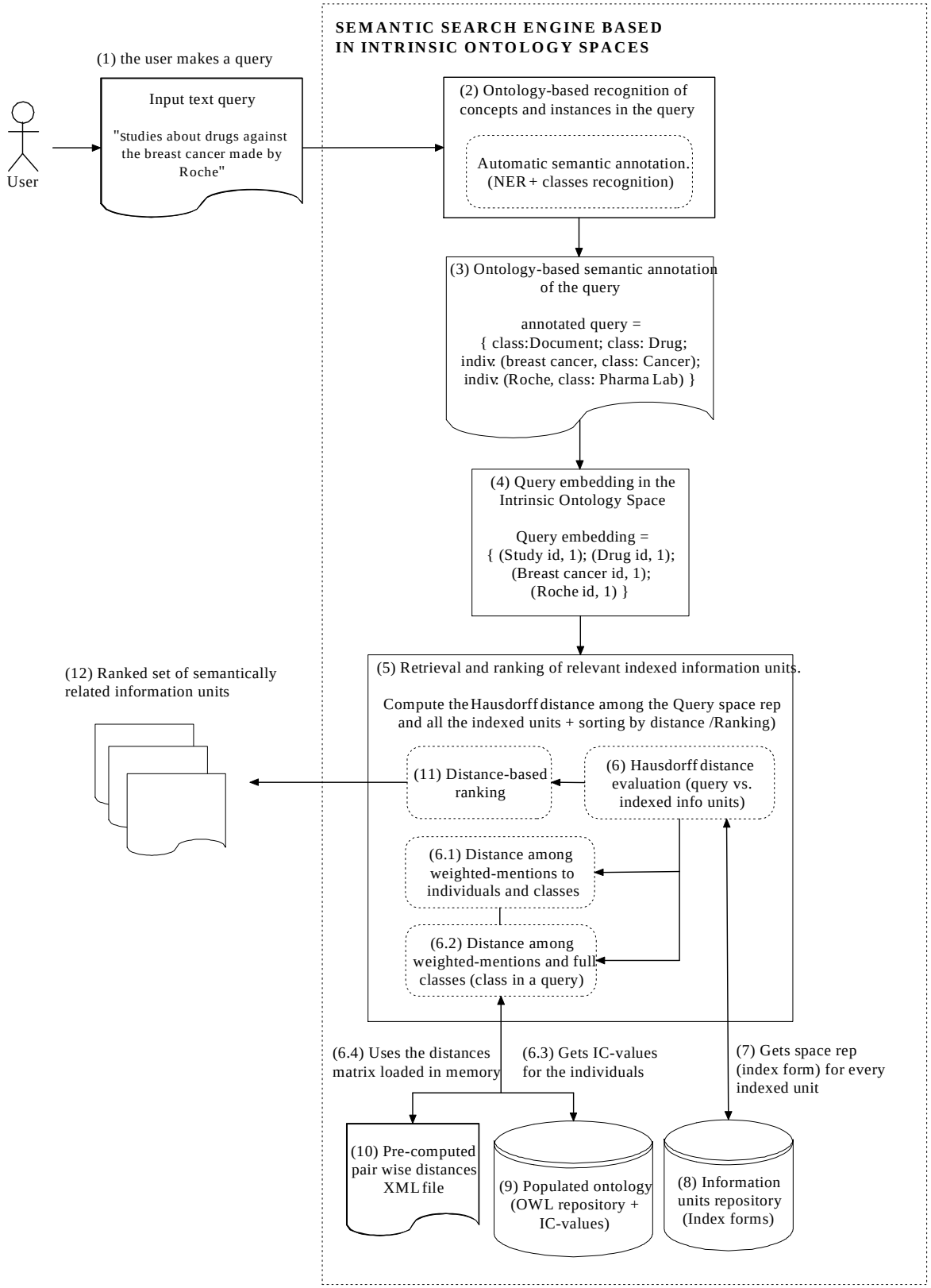


Figure 2.15: Retrieval and ranking process of indexed information units

ontology. The prior statement is not true when the ontology is a general poset encoding any multiple inheritance relation.

A hierarchical Voronoi diagram is a tree-like spatial partition where each node represent a subset of the global space, and its children define a Voronoi partition. Each node of the HVD graph defines a site for the local Voronoi diagram associated to its parent node in the tree. In most of the applications, the Voronoi cells are compact subsets of the global, by contrast, in our representation space, these cells will be open subsets.

The HVD structure is very useful for any task requiring a spatial partition, such as the search of near neighbors in clustering, classification or information retrieval. This structure has been used in image retrieval [Swets & Weng, 1999] and geographic information systems [Gold & Angel, 2006].

2.4.11 Summary and proof of the model

The *Intrinsic Ontological Spaces* are defined by the metric space (X, d_X), where X is the space of weighted-mentions to individuals and classes within a populated ontology $(C \cup I_C, \leq_C)$. The underlying representation space for annotated information units is defined by the metric space $(\mathcal{D}, d_{\mathcal{D}})$, where $d_{\mathcal{D}}$ is given by (2.34).

Note that $\mathcal{D}$ is the space for the frequency-based annotated input units (documents) and the distance $d_{\mathcal{D}}$ is the ontology-based metric induced by our model over the space of input information units. In the tables below we offer a summary with the definitions of the mathematical objects and functions included in the definition of the IR model proposed in this work.

Next, we prove that space (X, d_X) and the embedding functions for documents, individuals and whole classes, denoted by φ, φ_I and φ_C, fulfill the design axioms proposed in section 2.4.2, thus, the proposed ontology-based IR model is a structure-preserving representation space for any sort of semantically annotated data, that mimics the Jiang-Conrath distance for tree-like populated ontologies, including classes and individuals. Moreover, the distance among subsets of the representation, given by the metric $d_{\mathcal{D}}$, is a consistent extension of the ontology-based semantic ditstance for classes to individuals and collection of classes or individuals. In summary, we prove that the model is well defined from an algebraic point of view.

Theorem 1 (structure-preserving representation) *Given an ontology $(C \cup I_C, \leq_C)$, its associated* metric ontology *$\mathcal{O} = (C \cup I_C, \leq_C, d_{wJC})$, and a space of frequency-based annotations $D = C \cup I_C \times \mathbb{N}$. The functions pair (φ_I, φ_C) defined below is an Intrinsic Ontology Embedding, as well as the metric space (X, d_X) defined below is an Intrinsic Ontological Space, wherein $X = C \cup I_C \times [0, 1] \subset \mathbb{R}$.*

$$\begin{cases} \varphi_I : D \to X \\ \varphi_I\left(\tau_j, f_j^k\right) = \begin{cases} (\tau_j, 1), & \forall \tau_j \in I_C \\ (x_{C_i}, 1) & \forall \tau_j \in C \end{cases} \\ \varphi_C : C \to X \\ \varphi_C\left(C_i\right) = \{x \in X \mid \pi_O\left(x\right) \leq_C C_i\} \end{cases}$$

$$
\begin{aligned}
LA &: C \cup I_C \rightarrow \mathcal{P}(C) \\
LA(x) &= \begin{cases} x, & \text{if } x \in C \\ \{c_i \in C \mid x \leq c_i, \nexists c_j \neq c_i \text{ y } x \leq c_j \leq c_i\} \end{cases} \\
d_X &: X \times X \rightarrow \mathbb{R} \\
d_X((x, w_x), (y, w_y)) &= \begin{cases} \min\limits_{LA(x) \times LA(y)} \left\{ \begin{array}{l} -log(w_x \cdot P(x|LA(x))) \\ -log(w_y \cdot P(y|LA(y))) \\ +d_{wJC}(LA(x), LA(y)) \end{array} \right\}, & \text{if } x \neq y \\ \left| log\left(\frac{w_x}{w_y}\right) \right|, & x = y \end{cases}
\end{aligned}
$$

Proof:

Axiom 1. The proof of the order invariance (axiom 1) is trivial and it follows directly from the definition of the mapping φ_C in (2.17).

1: given two concepts $C_1, C_2 \in C$, such that $C_1 \leq_C C_2$.
2: from the definition of φ_C above, we get:
$$\varphi_C(C_1) = \{x \in X \mid \pi_O(x) \leq_C C_1\} \quad \text{and}$$
$\varphi_C(C_2) = \{x \in X \mid \pi_O(x) \leq_C C_2\}$.
3: from 2: we get that $\forall x \in \varphi_C(C_1) \Rightarrow \pi_O(x) \leq_C C_1$
4: but using the 3: plus the premise 1: and the transitivity
$\forall x \in \varphi_C(C_1) \Rightarrow \pi_O(x) \leq_C C_1 \leq_C C_2 \Rightarrow \pi_O(x) \leq_C C_2$
5: finally from 4: and the definition of $\varphi_C(C_2)$ in 2: we prove the axiom 1.
$\forall x \in \varphi_C(C_1) \Rightarrow \pi_O(x) \leq_C C_2 \Rightarrow \varphi_C(C_1) \subset \varphi_C(C_2)$

Axiom 2. The proof of the metric invariance (axiom 2) follows from the definition of the distance d_X.

1: given any two concepts $C_1, C_2 \in C$.
2: their whole image in the representation space is $(C_1, 1)$ and $(C_2, 1)$.
3: replacing in $d_x((C_1, 1), (C_2, 1))$ we get
4: $d_X((C_1, 1), (C_2, 1)) = -log(1) - log(1) + IC(C_1) - IC(LA(C_1)) + IC(C_2) - IC(LA(C_2)) + d_{wJC}(LA(C_1), LA(C_2))$
5: but $LA(C_1) = C_1$ and $LA(C_2) = C_2$
6: thus from 5: and 4: we prove the axiom 2 $\rightarrow d_X((C_1, 1), (C_2, 1)) = d_{wJC}(C_1, C_2)$

Axiom 3. The inclusion invariance (axiom 3) follows from the definition of the mapping φ_C. According to the definition (2.17), the image $\varphi_C(C_i)$ of any class C_i subsumes the image $\varphi_I(\tau)$ of any individual $\tau \in C_i$, thus the axiom 3.a is verified. Moreover, by induction, the verification of the axiom 1 also implies the verification of the axiom 3.b, because the image $\varphi_C(C_j)$ of any class C_j also subsumes the image $\varphi_C(C_i)$ of any descendant class $C_i \leq C_j$, thus,$\varphi_C(C_j)$ also subsumes the images of any individual $\tau \in C_i$.$\boxdot$

Id	Element of the model	Definition and notation
1	Input base ontology (taxonomy)	$T = (C, \leq_C)$
2	Instances of any class	$I_C = \{I_{C_i}\}$
3	Populated base ontology	$O = (C \cup I_C, \leq_C)$
4	Metric ontology	$\mathcal{O} = (C \cup I_C, \leq_C, d_C)$
5	Frequency weighted-set space D of input information units	$D = C \cup I_C \times \mathbb{N}$
6	Input information unit $\delta_k \subset D$	$\delta_k = \left\{ \left(\tau_j, f_j^k\right) \in C \cup I_C \times \mathbb{N} \mid j \in J(k) \right\}$
7	Ontology representation space	(X, d_X), where $X = C \cup I_C \times [0,1] \subset \mathbb{R}$ and $d_X : X \times X \to \mathbb{R}$ is a metric on X
8	Intrinsic ontology embedding (structure-preserving)	A function pair $\Phi = (\varphi_I, \varphi_C)$ that verifies the axioms below: (1) $C_1 \leq_C C_2 \Rightarrow \varphi_C(C_1) \subset \varphi_C(C_2)$, $\forall (C_1, C_2) \in C \times C$ (2) $d_C(C_1, C_2) = d_H(\varphi_C(C_1), \varphi_C(C_2))$, $\forall (C_1, C_2) \in C \times C$ (3) $\varphi_I(\tau) \subset \varphi_C(C_i), \forall \tau \in I_{C_i}$ $\varphi_I(\tau) \subset \varphi_C(C_j), \forall \tau \in I_{C_i}, \forall C_j | C_i \leq_C C_j$
9	Intrinsic embedding φ_I for individuals	$\varphi_I : D \to X$ $\varphi_I\left(\tau_j, f_j^k\right) = \begin{cases} (\tau_j, 1), & \forall \tau_j \in I_C \\ (x_{C_i}, 1) & \forall \tau_j \in C \end{cases}$
10	Whole class embedding φ_C (classes in queries)	$\varphi_C : C \to X$ $\varphi_C(C_i) = \{x \in X \mid \pi_O(x) \leq_C C_i\}$
11	Info units embedding φ and static weigths ω_j^k (indexes)	$\varphi_{\mathcal{D}} : \mathcal{D} \to X$ $\varphi_{\mathcal{D}}(\delta_k) = \left\{ \left(\tau_j, \omega_j^k\right) \in X \mid j \in J(k) \right\}$ $\omega_j^k = \frac{f_j^k}{\sum_{j \in J(k)} f_j^k}$

Table 2.4: Summary of the objects defined in the Intrinsic Ontology Spaces model

Id	Elements of the model	Definition and notation
12	Semantic weighting $\widehat{IC}$ (normalized IC values)	$\widehat{IC} : \delta \subset D \rightarrow \mathbb{R}$ $\widehat{IC}(\delta_k) = \sum_{j \in J(k)} \underbrace{\omega_j^k}_{\text{static weight}} \cdot \underbrace{IC(\tau_j)}_{\text{dynamic \& semantic}}$,
13	Weighted Jiang-Conrath distance d_{wJC}.	$d_{wJC} : O \times O \rightarrow \mathbb{R}$ $d_{wJC}(a,b) = \min_{x \in P(a,b)} \left\{ \sum_{e_{ij} \in x} w(e_{ij}) \right\}$ $w : E \rightarrow \mathbb{R}$ $w(e_{ij}) = IC(P(v_i \vert v_j)) = -log_2 P(v_i \vert v_j)$
14	Metric of the IOS representation space	$d_X : X \times X \rightarrow \mathbb{R}$ $d_X((x,w_x),(y,w_y)) =$ $\begin{cases} \min_{LA(x) \times LA(y)} \left\{ \begin{array}{l} \underbrace{-log(w_x \cdot P(x \vert LA(x)))}_{(1)} \\ \underbrace{-log(w_y \cdot P(y \vert LA(y)))}_{(2)} \\ \underbrace{+d_{wJC}(LA(x), LA(y))}_{(3)} \end{array} \right\} & \text{, if } x \neq y \\ \left\vert log\left(\frac{w_x}{w_y}\right) \right\vert & \text{, } x = y \end{cases}$
15	Intrinsic Ontology Spaces	(X, d_X), with d_X defined in (14) above
16	Hausdorff distance	$d_H : A \in \mathcal{P}(X) \times B \in \mathcal{P}(X)$ $d_H(a,b) = max\left\{ \sup_{a \in A} \{d_X(a,B)\}, \sup_{b \in B} \{d_X(b,A)\} \right\}$
17	Ontology-based ranking d_D	$d_{\mathcal{D}} : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}$ $d_{\mathcal{D}}(\delta_k, \delta_m) = d_H(\varphi(\delta_k), \varphi(\delta_m))$
18	Distance in the representation space between any weighted-mention and the image of a whole class (query mention), denoted by d_{IC}.	$d_{IC_i} : X \times \varphi_{C(C_i)} \subset X \rightarrow \mathbb{R}$ $d_{IC_i}\left(\left(\tau_j, \omega_j^k\right), \varphi_C(C_i)\right) =$ $= \begin{cases} 0, & \text{if } \tau_j \leq_C C_i \\ d_X\left(\left(\tau_j, \omega_j^k\right), (x_{c_i}, 1)\right), & \text{other case} \end{cases}$

Table 2.5: Summary of the objects defined in the Intrinsic Ontology Spaces model

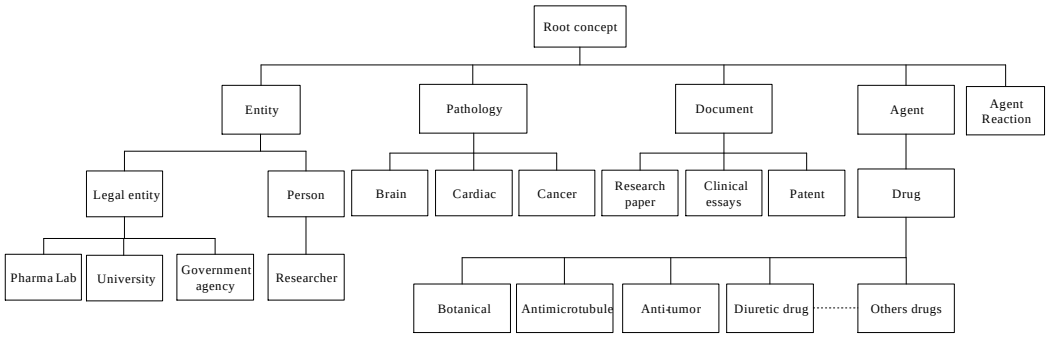


Figure 2.16: A small ontology for the indexing of bioengineering documents

2.5 Example of use

In this section, we provide one toy example to show the operation of the proposed model, however, we are plenty aware that it is necessary to carry-out in a near future an exhaustive set of experiments to compare our model with other ontology-based and classic IR models.

Nowadays, the evaluation of ontology-based IR models is an open problem due to the lack of standard benchmark corpus specifically developed for this task. A first try in this trend is the work described in [Fernández et al., 2009]. In this work, the authors introduce a benchmark based in TREC data which comprises: (1) a corpus of text document, (2) a set of queries with their corresponding document relevance judgments, and (3) a set of ontologies covering the query topics.

Although the corpus developed by the Fernández et al. could be a good starting point for a large scale evaluation of our model, a serious drawback in our case is the lack of the semantic annotation data. The cornerstone of our model is the metrization of the representation space using an ontology-based semantic distance, by this reason, the right approach to compare our model with other ontology-based IR models would be to start from the same semantic annotation data, with the aim of removing this variable from the experiments. If the semantic annotation data is not fixed in the benchmarks, the results will depend on the performance of the automatic semantic annotation components used by each IR system. Such as you can appreciate, the search and selection of an proper corpus to carried-out the validation experiments of the proposed IR model is not an easy task, and it is our next priority task in the short term.

Next, we show a toy example to explain the operation of the proposed IR model. In figure 2.16, we introduce an ontology example in the bioengineering domain which is used as base ontology in the figures 2.14 and 2.15. The ontology defines some types of documents to be indexed, as well as some types of entities, pathologies and sort of agents to be recognized within the indexed papers. The aim of the ontology is to organize the content of the indexed documents and for supporting the search and retrieval process. The application is an indexing system for research papers, clinical essays and patents in the bioengineering field.

In figure 2.14 and 2.15 we can appreciate, respectively, the indexing and retrieval processes based in the proposed IR model. Some user provides a research paper as input document, and an automatic semantic annotator identifies the mentions

to individuals and classes defined within the base ontology above. The semantic annotations are used to compute the static representation of the document, as a set of tuples of weighted-mentions to ontology objects. The tuples define the index form for the information units, which is stored in the indexes repository. By last, the weighted annotations are inserted within the populated base ontology, including the cross-references to the source indexed units.

2.6 Expected problems

Up to now, we have introduced a novel ontology-based IR model, and we have proven that our model fulfills a set of principle-based axioms to preserve the intrinsic geometry of the base ontology, and that the model is well founded. Nevertheless, the model has not even evaluated experimentally, thus, it is likely that the model could need to be modified in some way while the structure-preserving approach is maintained.

By example, one possible problem that we could find is that the Hausdorff distance among documents and queries is dominated by the farthest elements among subsets, according to the semantic distance. As consequence, one document that semantically matches the query could be moved to distant positions in the ranking if it contains some annotation far away from the concepts or entities in the query or vice versa. This drawback, as well as others unknown now, will be studied and managed during the next experimental stage.

2.7 Conclusions

In this work, we have introduced a novel ontology-based IR model based in a structure-preserving approach, inspired by a geometric point of view. Moreover, we have also introduced a novel ontology-based semantic distance, called *weighted Jiang-Conrath distance*, which is based in a generalization of the standard Jiang-Conrath distance to any sort of taxonomy.

The *weighted Jiang-Conrath distance* is defined as the shortest path on the weighted-graph associated to the base ontology, whose edge weights matches the Jiang-Conrath distance for tree-like taxonomies. The novel distance solves the two main drawbacks of the standard Jiang-Conrath distance as follows: first, the novel distance is uniquely defined on any general taxonomy, and second, it is a metric on any sort of taxonomy.

The main idea of the novel IR model is to build an embedding of any populated ontology into a metric space of weighted-mentions, while the intrinsic geometry of the ontology is preserved. The model has been described and justified from a theoretical point of view, but, it still being necessary to evaluate it experimentally, to prove our main hypothesis about the expected improvements in the ranking quality, and the precision and recall measures, as main result of the structure-preserving approach of the model.

The proposed model integrates in a natural way the intrinsic geometry of the any ontology, which is defined by three algebraic structures: (1) its poset structure, (2) its metric structure derived from any ontology-based semantic distance, such

as the Jiang-Conrath distance used in our model, and (3) the set and subsumption relations among the classes and individuals of a populated ontology. We have proved that the Intrinsic Ontological Spaces are well defined metric spaces on any sort of ontology with general poset structure.

Up to our knowledge, the proposed model is the first one to be completely ontology-based and structure-preserving, in the sense that all their components are ontology-based, such as the embedding, ranking. weighting, indexing, retrieval and storing processes. Some features of the model make possible to remove the necessity to invoke any SPARQL engine to retrieve the related documents by integrating retrieval, ranking and weighting of the populated ontology in the same model, although it is not an exclusive feature of the model.

The proposed model solves the modeling gaps identified in prior ontology-based IR models reported in the literature. First, the orthogonality constraint imposed by all the vector space models is removed through the integration of a real intrinsic semantic distance derived from the ontology. Second, the cardinality mismatch induced by mixing sets and elements at the same representation level is removed through the specific integration of these relations in the model. Third, the ranking methods defined by the cosine function among vectors whose spatial relations are statistical, not semantic, are substituted by the Hausdorff distance among sets of classes and individuals according to the intrinsic semantic distance on the ontology. Moreover, the use of the Hausdorff distance solves some continuity problems reported in the pioneer work of [Rada et al., 1989].

The representation space for the documents, or information units, is a metric space whose structure and distance function mimics the ontology-based semantic distance among the concepts of the ontology, therefore, the model integrates all the semantic knowledge encoded in the taxonomic structure of the ontology.

Chapter 3

A novel manifold-based text classifier

This chapter introduces a novel method for text categorization, problem known as TC by its acronym in English language. The novel method can be categorized as a *manifold-Bayes* hybrid model, and we call it *Intrinsic Bayes-Voronoi classifier.* The core of the method is the building of a geometric representation of the classes which uses the intrinsic geometry of the *features space*, wherein the features space is defined by the *positive unit hypersphere*. The geometry of the classes is represented by the normal distribution of a random vector on the hypersphere. For this, we introduce the concept called *geodesic normal distribution* $\mathcal{N}_g(\mu, \Sigma_g)$ as a generalization of the Gaussian normal distribution to domains defined by differential manifolds. The *geodesic normal distribution* is built as a function of the geodesic distances and the tangent space of the underlying features space, and it induces the definition of *geodesic Mahalanobis distance*. This intrinsic distribution allows us to build an optimal Bayes classifier based on the intrinsic geometry of the data and the features space. Also, we introduce a distance function on the features space that we call Bayes distance, which allows us to demonstrate that the intrinsic Bayes classifier defines an intrinsic Voronoi diagram on the features space, relationship that gives name to the method.

3.1 Introduction

The growth of the Web and the proliferation of large collections of documents in companies and government agencies, has motivated the development of document classification algorithms for their indexing, storing and retrieval. The importance of this research topic is highlighted in the review in [Sebastiani, 2002b], where the author notes that TC techniques are being used in many applications, such as document indexing based on a vocabulary, spam filtering, sense disambiguation, and the automatic classification of web pages, among others.

The single-label document classification involves assigning each unseen document to a subject category, while the multi-label document classification admits to assign one document to more than one class. According to the definition of the classifier, the problem reduces to find an approximation function $\widetilde{\varphi}$ to the unknown classification

function φ, such that the error probability be minimized.

Definition 21 (Classifier) *Given a set of classes C and a set of documents D, where the documents are defined by vectors according to the VSM model, a classifier φ is a binary function $\varphi : D \times C \rightarrow \{T, F\}$.*

Most of document classification methods are based on the VSM model, known as "bag of words", whose main properties are described in section 3.3.1.

The TC problem has been broadly investigated during the last two decades, such as is witnessed by the revisions of Yang and Liu [Yang & Liu, 1999], and Sebastiani [Sebastiani, 2002b] [Sebastiani, 2005]. Throughout this period, all sort of machine learning algorithms have been proposed for its solution, such as:

- Information retrieval methods [Rocchio, 1971].
- Regression models [Yang & Chute, 1994], [Fuhr et al., 1991].
- Nearest neighbors classification (kNN) [Masand et al., 1992], [Yang, 1994], [Yang, 1999] and [Lam & Ho, 1998].
- Bayesian methods [Tzeras & Hartmann, 1993], [Lewis & Ringuette, 1994], [Moulinier, 1997], [Koller & Sahami, 1997], [McCallum & Nigam, 1998], [Baker & McCallum, 1998] and [Chai et al., 2002].
- Decision trees: [Fuhr et al., 1991], [Lewis & Ringuette, 1994] and [Moulinier, 1997].
- Rule-based learning [Apté et al., 1998], [Cohen, 1995], [Cohen & Singer, 1999] and [Moulinier et al., 1996].
- Neural networks [Wiener et al., 1995] and [Ng et al., 1997].
- Support Vector Machines (SVM) [Joachims, 1998], [Kwok, 1998].

Recently, a novel family of TC methods called manifold-based has emerged in the literature, whose main feature is the encoding of the geometry of the problem in the model. Among these pioneering works, we can cite the contributions in [Zhang et al., 2005] and [Cai & He, 2012]. Our research in this chapter, as well as the classification method proposed, are framed in this family of manifold-based methods, but its approach can be best categorized as an hybrid method of type *manifold-Bayes*, because it implements a Bayesian inference on a differential manifold.

According to different reports in the literature, such as [Sebastiani, 2005] and [Lewis et al., 2004], the keyword-based TC method with best results and wider acceptance is the SVM-based method [Joachims, 1998], which is introduced by Joachims in one of the most cited works in the field.

Due to the maturity of the current TC methods, and the broadly acceptance of SVM [Joachims, 1998], most of the research effort has focused in the solution of other related problems, such as: (1) the selection of features before the application of any machine learning algorithm [Dasgupta et al., 2007]; (2) the use of semantic

features [Chua & Kulathuramaiyer, 2004] [Khan et al., 2010]; (3) the active learning for the selection of the most discriminant features [Esuli & Sebastiani, 2009]; (4) unsupervised learning methods (clustering) [Sandler, 2005]; and (5) hierarchical classification [Esuli et al., 2008]. Despite of this shift in the research trends, we continue in this work with the exploration of novel classification and machine learning methods for text categorization, following a geometry-based approach.

Browsing the literature, we can identify the following list of problems and limitations related with the text categorization problem:

(p1) *Corpus dimensionality.* The size of private document collections and the Web has grown exponentially, with a range from 10^3 to 10^{10} indexed documents, which sets a huge challenge for any classification method.

(p2) *Dimensionality of the features space.* Due to the size of the corpus and the number of classes, it is common that the vocabulary size is in the range from 10^4 to 10^6 terms. Citing an old example, the RCV2 corpus [Lewis et al., 2004] uses vectors with a dimension close to 50000.

(p3) *Feature selection.* The methods for feature selection have become a necessity, as well as an active research line in the field of text categorization, such as is noted in [Dasgupta et al., 2007]. The aim of these methods is to allow a better discriminative ability while the computational cost of the classifiers is reduced as consequence of the reduction of the dimensionality of the features space. The feature selection is a hard problem and it is independent of the used features, by this reason, the classification methods that do not require to use feature selection, such as VSM, have become very popular.

(p4) *Black-box use of machine learning methods.* Most of the proposed methods have been adaptations, or reuses of general-purpose machine learning algorithms, mostly using libraries as black-box components, such as LibSVM[1] [Chang & Lin, 2011]. This statement is true, including in the case of well established methods as SVM [Joachims, 1998].

(p5) *Non-linear methods.* The classification methods that define non-linear decision boundaries have some drawbacks derived from the high dimensionality of the features spaces, and they use often some sort of data transformation, such as the kernel trick, with the aim to reduce the classification to a linear model in the space of transformed features. The definition of these transformations and their effect on the classification results is not very clear in most of cases.

(p6) *Geometric representation of the problem.* Most of methods, including SVM, are working on the Euclidean space in an explicit or implicit way, although there are some exceptions in the family of *manifold-based* methods. The lack of a precise representation of the intrinsic geometry of the features space and the classes, introduces some alterations in the model, which are often unseen, since most of time the models are using a projection of the real data model.

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

(p7) *Probability model of the data.* The Bayesian methods have been very popular by its simplicity and easily to get acceptable results, however, these models exhibit some drawbacks due to their implicit assumptions. Usually, the Bayesian methods represent the data as a mixture of Gaussian or multinomial distributions in the Euclidean space, assuming in most of cases that the coordinates of the vectors are independent random variables, and that all the classes share the same distribution. These hypothesis are far to be realistic. In this work, we represent the classes as a mixture of random vectors with different Gaussian distributions, which are defined on the geometry of the features space. The proposed distributions use the full model of dependencies among vector coordinates, thus they model better the nature of the data that other simplified models.

(p8) *Optimal solution characterization and features space partition.* About this point, we will study the conditions on the existence and unicity of the proposed model, and the geometric structure induced by the solution on the features space.

(p9) *Scalability of the number of classes.* Any industrial automatic classifier of documents should be able to support the progressive addition of novel document categories without affecting the operation of the system, neither obligate to any subsequent configuration, training, or additional development.

3.1.1 Our method

The proposed method is called *Intrinsic Bayes-Voronoi Classifier*, whose acronym is *IntBayesVor.* The main novelty of the method is the definition of an optimal Bayes classifier that represents the geometry of the classes of documents by a *geodesic normal distribution* on the features space defined by the *positive unit hypersphere.* Because the document vectors are normalized to have unit norm, the vectors are contained in a subset of the Euclidean space, defined by the unit hypersphere.

For the definition of the normal distribution for the classes, we generalize the standard multivariate normal distribution defined on the Euclidean space, to a normal distribution on a differential manifold. We extract the normal distribution on a unit hypersphere as a particular case with a closed formula. In this way, we are modeling the classes as a mixture of normal distributions on the features space, which we denote by [2] $\mathcal{N}_g(\mu, \Sigma_g)$. In this generalization, we also identify and define the *geodesic Mahalanobis distance*, which is a generalization to differential manifolds of the measure with the same name.

Using the geodesic normal distribution, we define a distance function that we call *intrinsic Bayes distance*, which induces a partition on the features space that we call *Intrinsic Bayes-Voronoi diagram.* This diagram allows us to derive our multiclass classification method called *IntBayesVor.* This classifier is simply an optimal Bayes classifier on the features space considered as a differential manifold. The resulting

[2] We will use the letter "g" as subindex, or function name, in reference to "geodesic", to denote mathematical objects that are defined by the intrinsic geometry of the features space, considered as a differential manifold.

Voronoi diagram is an abstract generalization of the classic Voronoi diagrams for differential manifolds, through the definition of a particular local distance function on the manifold.

Following the analysis, we introduce a theorem to prove that under the hypothesis of a geodesic normal distribution $\mathcal{N}_g$ for the classes, the optimal Bayes classifier defines an *Intrinsic Bayes-Voronoi diagram.*

Once the intrinsic Voronoi diagrams are built, a relationship is established between the distance functions derived from the distributions, and the geometry of the decision boundaries, what characterizes the geometry of the decision regions.

The Bayes distance unifies in a same mathematical object the geometry of the features space and the geometry of the data defined on it, while it also defines a mapping from the set of geodesic normal distributions in the space of Voronoi diagrams, by selecting the diagram which maximizes the probability a posteriori that the document seen belongs to a given class.

The method introduces a framework where other distributions, different from the normal, could be defined as intrinsic distributions on a differential manifold, leading to the possibility to use other probability models.

Once the method *IntBayesVor* has been introduced, we show that its computation is not possible due to the high dimensionality of the features space. By this reason, we introduce a practical method to approximate the optimal solution through the use of the χ^2 feature selection method, and PCA [Zu et al., 2003] for the approximation of the covariance matrix. Here it should be noted that we use PCA to approximate the covariance matrix of the data for each class, not to build a LSI representation with the more representative terms for all the classes at the same time. The features filtering is made by the χ^2 method before the PCA decomposition, task that could be unnecesary if the set of non-null components of the features vector does not exceed a predefined threshold of maximum dimension.

The described novelties allow us to make some contributions in some points described in the previous section as follows:

(p4) Unlike other methods based in general-purpose algorithms, our method makes an exhaustive use of the specific properties of the TC problem with the aim to improve the performance of the system.

(p5) Our method is intrinsically non-linear, because the decision boundaries can have any arbitrary shape according to the distribution of the data vectors on the geometry of the features space.

(p6) The proper geometric representation of the features space and the data are the core idea of our method. Our main hope is to get a better representation for the data, and as consequence, a classifier with better performance.

(p7) Like generative model for the data, we propose a Gaussian mixture with independent distributions for each class, where each distribution is defined on the geometry of the tangent space to the features space. In spite that the multinomial distribution has got a broadly acceptance, we expect that the proposed model can improve the representation of the data as a consequence of the use of the geometry of the features space.

(p2, p3) To provide a practical method that can work on any high dimensionality features space, we introduce an approximated solution to the proposed model, which uses a feature selection method (χ^2) and other one of dimensionality reduction (PCA). These methods allow us to define the intrinsic normal models on specific subspaces for each class. Like natural evolution of the proposed model, we could evaluate other feature selection methods without to affect the general framework of the classification method.

(p9) The definition of generative models on subspaces for each class also allows to provide a practical solution with regard to the scalability in the number of classes, because the set of classes can be extended in a simple way, through the computation and storing of the model for each new class, without the need to modify the existent vocabulary, or the generative models of the known classes.

3.1.2 Structure of the chapter

In section 3.2, we review the related work with our investigation. In section 3.3, we introduce some preliminary concepts to make easier the reading of the work, and we describe the intrinsic geometry of the features space for the TC problem. The section 3.4 constitutes the core of this chapter, and it is the place where our method for document classification is introduced. In section 3.5, we describe the experiments carried-out and the results obtained. By last, we summarize our conclusions in section 3.6.

3.2 Related work

Our work is related with the Bayesian-type and manifold-based text classification (TC) methods. Due to the large number of revised works, we have divided the section in two parts, one for each family. We have also included a brief introduction about the most used probability models to enlighten some of the main features exhibited by both families of methods.

3.2.1 Bayesian methods

Despite of their drawbacks, the Bayes classifiers have enjoyed wide acceptance in various fields. This fact is largely due to its simplicity, efficiency, and good theoretical foundation.

The Bayesian methods use a probability distribution model for each class, whose parameters are estimated through a training dataset. Later, the classification of a new document is made through the application of the Bayes rule: the document gets the label of the class which maximizes the a posteriori probability. Most of these Bayesian methods have used binomial or multinomial distributions, and only in some cases, the Gaussian normal distribution has been used.

In [McCallum & Nigam, 1998], the authors make a survey of the text categorization methods of type "Naive Bayes". The methods are subdivided in two categories according to the probability model used: binomial or multinomial. The binomial model only assigns binary values for each component (feature) of the document

vectors, thus these methods do not consider the frequency of the features within a document. The binomial models represent every document by a binary vector, and its occurrence probability is the product of the occurrence probabilities for each feature, given the specific distribution of each feature according to the class. In contrast, the multinomial model considers the frequency for the features within a document, thus, each document is represented by a vector of whole numbers. Like the binomial case, the document probability is the product of the probabilities for each feature (coordinate). The authors conclude that the multinomial model offers better results that the binomial one.

In [Koller & Sahami, 1997], the authors introduce a document classifier based in a Bayesian network. The network represents the computation of the a posteriori probability for a class, given a features vector. To narrow the network complexity, they propose as first step the use a feature selection method to reduce the dimensionality of the problem. Later, a Bayesian network is defined where each variable (node) has, at most, one or two parent nodes (dependencies). The classifiers for each class are defined in such way that every classifier only uses a set of relevant features per class. The method does not consider the geometry of the features space, and it uses a general-purpose classification method, the KDB [Sahami, 1996]. The Bayesian networks are too restrictive because they limit the number of possible dependencies among variables, with independence of the dimension of the features space.

In [Lewis, 1998], the author makes a survey about the application of the Naive Bayes classifier to the TC problem, and analyzes the different vector models used in the literature: binomial, multinomial and normalized. The method represents the classes by a probability distribution, assuming that the components of the document vector are independent variables (Naive Bayes hypothesis). The classification is made using the Bayes rule.

In [Stamatatos et al., 2000], the authors propose a method for the classification of documents according to their genre and author, which is based in the use of a Gaussian normal mixture and the Mahalanobis distance in the Euclidean space. The work reports comparable results to other methods reported in the literature, while it also endorses the use of a Gaussian distribution to represent documents.

Torkkola [Torkkola, 2004] proposes a method for document classification that uses three steps: a dimensionality reduction based in LSI, a data transformation through the linear discriminant analysis technique (LDA), and by last, a linear SVM classifier. The proposed classifier assigns a single category per document. LDA is a dimensionality reduction method [3] which builds a transformation matrix to project the data on a space with lower dimension. The transformation matrix is defined as a function of two covariance matrixes which measure the intra-class cohesion and the inter-class separation [4]. The LDA method assumes that the classes are generated by a mixture of normal distributions in the Euclidean space, where all the classes share the same distribution, implying that the decision boundaries are linear. Due to the computational complexity for the whole LDA method, Torkkola proposes a

[3] LDA also refers to the linear discriminants of Fisher for Gaussian mixtures with the same variance.

[4] This statistical concepts are equivalent to the optimization measures used by the *clustering* methods.

practical method which uses as first step a dimensionality reduction through the known LSI method, therefore, the method works in three steps: LSI-LDA-SVM.

In [Schneider, 2005], the author introduces some proposals to improve the performance of the methods for document classification through the use of basic Bayes classifiers. The classifiers of type "naive" Bayes assume that the document vector is composed by a set of independent variables, and it implies that the covariance matrix is diagonal. Schneider propose a multinomial model to represent the data distribution in the Euclidean space. Some of the proposed improvements are as follows: (1) to apply a logarithm transformation to the vector components to reduce the dominant effect of the large values; (2) to use a feature selection method, such as the one described in [McCallum & Nigam, 1998], which use the *mutual information* index, or the method proposed by Schneider in his work, called *Cluster Representation Quality*; (3) to assume that all the a priori probabilities are equal to remove their dominant effect in some cases.

In [Genkin et al., 2007], the authors introduce a method for document classification of Bayesian type, which is based in the known statistical technique, called *logistic regression (LR)*. The novelties of their work are focused in the adaptation of the standard LR method to the TC problem, solving the underlying dimensionality problem for the text categorization problem. The core idea of their method is to use a probability model which induces sparse matrixes in the resulting model. The sparsity of the matrixes is exploited by the method to reduce the computational cost of the whole system. The LR model defines the a posteriori probability of the classes as a radial basis function (logistic function (3.2) for a linear combination β (3.1) of the components in a data vector on the Euclidean space, in this last feature, the method is similar to the perceptron.

$$p\left(c_x \mid \beta, x_i\right) = \psi\left(\beta^T x_i\right) = \psi\left(\sum_j \beta_j x_i^j\right) \tag{3.1}$$

$$\psi\left(r\right) = \frac{e^r}{1+e^r} \tag{3.2}$$

The linear combination $\beta^T x_i$ represents a linear decision boundary on the Euclidean space, defined by the vector β, while the logistic function ψ acts as a smooth decision threshold which adds a non-linear component to the model. The training of the LR model assumes the estimation of the decision boundary β for the distribution of each class. In [Genkin et al., 2007], the authors assume an equal univariate normal distribution for all the components β^j of the parameters vector, a hypothesis not very plausible. Unlike our approach, the LR distribution does not represent the random behavior of the vectors using the geometry of the features space, in contrast, they represent the vector components β that define the decision boundary. LR can also be interpreted like a feature selection method of probabilistic type, because the parameters β are weighting the contribution from each feature to the a posteriori probability $p\left(c_k|x\right)$. The proposed method improves the results provided by other methods, such as the linear SVM model and the Ridge Logistic Regression model.

In [Mouratis & Kotsiantis, 2009], Mouratis and Kotsiantis propose an improvement in a TC method based in the Bayesian discriminant DFE. The DFE model is

a general-purpose classifier developed by Su et al. [Su et al., 2008]. The proposed improvement consists in the use of feature selection method based in the known chi-square function χ^2 plus the representation of the data by a multinomial distribution. Using these simple improvements, the authors achieve to improve the results of the linear SVM method when the method is evaluated on the corpus RCV1-v2 [Lewis et al., 2004].

In [Feng et al., 2012], the authors investigate the TC problem for documents in Chinese language. Feng et al. propose a feature selection method specific for each class, based in the use of a different generative probabilistic model for each one. Given an initial number of features d, matching the cardinal of the vocabulary of the VSM model, they build a real-valued function on the space of admissible selections (feature subsets) with signature $f_{c_i} : 2^d \rightarrow \mathbb{R}^+$, where $f_{c_i}(\Delta)$ is the conditional probability to watch the whole collection of documents assigned to the class c_i, for each admissible selection Δ. In this way, the feature selection problem is transformed in one optimization problem: the search of the optimal selection Δ^*, problem that is solved using a stochastic search algorithm. The categories are represented by a mixture of binomial distributions, under the hypothesis that the components of the document vectors are independent (Naive Bayes hypothesis).

3.2.2 Manifold-based methods

The manifold-based methods are characterized by trying to integrate in its model the intrinsic geometry of the data or the associated distribution. We consider that this novel family of classification methods represents the most innovative methods recently published about the TC problem.

Many of the manifold-based ideas are emerging in different application fields. For example, in the field of image processing we can cite the works in [Yan et al., 2007], [Arandjelovic et al., 2005], [Wang & Chen, 2009] and [Song & Tao, 2010].

Here, we make a survey of the methods for document classification and image classification that are related with our research. We have divided the manifold-based methods in two subcategories: (1) the methods that represent the geometry of the data using the distributions space, called *statistics-manifold* methods, and (2) those that use the features space to represent the geometry of the data, called *data-manifold* methods, among which is included the method proposed in this chapter.

3.2.2.1 Geometry of the distributions space: statistics manifolds

In [Lebanon, 2005], Lebanon notes that the classical text classification methods represent the documents as vectors in the Euclidean space, because most of them use general-purpose machine learning algorithms, such as we note in the introduction (see p4) to this chapter. His method uses the differential manifold structure [do Carmo, 1992] on a distributions space $\mathcal{P}$, defined by a set of parameters Θ, which defines an object called *statistical differential manifold*. The pair (Θ, g) represents a Riemannian manifold (smooth metric space), where Θ is the parametrization of the distributions space $\mathcal{P}$, and g is a Riemannian metric called *Fisher information metric* [Amari et al., 1987]. The work also introduces an embedding method of the data into the multinomial distributions space, what is equivalent to map every doc-

ument vector to the estimated distribution that explains better the seen data. The embedding function is defined for the multinomial distribution by exploiting the fact that the VSM model with a normalized TF-L_1 type, can be directly used to estimate the multinomial distribution, but unfortunately, this approach can not be extended to their distributions like the Gaussian one. Finally, the distance among points (x, x') in the features space is transformed to the geodesic distance among distributions $\left(\widehat{\theta}, \widehat{\theta}'\right)$ on the statistical manifold. Lebanon uses his novel distance like a kernel for some standard methods which are modified to include the new metric, such as kNN, SVM y LR. Most of ideas in this work already appear in a previous work of the same authors [Lebanon & Lafferty, 2004], although in [Lebanon, 2005], the model is explained better as well as its integration with the methods already cited. The experimental results are included in Lebanon's thesis [Lebanon, 2006b].

In [Lebanon, 2006a], the author propose to learn a Riemannian metric on the distributions space considered like a differential manifold. The metric is selected from a parametric family subject to the criterion of maximizing the inverse volume data set, measured on the statistical manifold. Once the metric is applied to the TC problem, the resulting geodesic distance is similar, but with better results, to the cosine similarity function, suggesting its use as a distance measure for TC algorithms.

In [Zhang et al., 2005], the authors introduce a method called *Multinomial Manifold,* which represents the intrinsic geometric structure of the data associated to a given class, and is defined on the multinomial distributions space $\mathcal{P}$. The main contributions are a theoretical proof about the plausibility of the geodesic distance on the distributions space like kernel for the SVM method, and the experimental results of these ideas on some corpus. The proposed method derives from results prior published by Lebanon. The distributions space is endowed with a differential structure through the parametrization $\Theta : P^{n+1} \rightarrow \mathcal{P}$, where the function's domain is $P^{n+1} = \left\{\theta \in \mathbb{R}^{n+1} : \sum_{i}^{n+1} \theta_i = 1; \forall \theta_i \geq 0\right\}$ [5]. Each multinomial distribution P_θ is described by a parameters vector $\theta = (\theta_1, \ldots, \theta_{n+1})$, such that $\Theta(\theta) = P_\theta$, where each element of the domain Θ selects a specific distribution. For the multinomial distributions, the values θ_i designates the occurrence probability for a number of instances of a feature w_i included in the vocabulary of the associated VSM model. The authors use a result in statistical differential geometry [Kass, 1989] which sets that the space of parameters P^{n+1} is isometric to the positive hypersphere S_+^{n+1} with *radius* equal to *2*. Using this property, they build an embedding function $\varphi : P^{n+1} \rightarrow S_+^{n+1}$ [Kass, 1989] which maps every distribution P_θ to a point on the hypersphere. The authors develop a kernel-type metric $k(\theta, \theta')$ where the distance is measured on the space of distributions on the hypersphere, using the geodesic arc-length among distributions. In this way, each document d is mapped to a distribution $P_{\widehat{\theta}_d}$ on the multinomial distributions space, while the kernel $k\left(\widehat{\theta}_j, \widehat{\theta}_d\right)$ computes all the geodesic distances among pairs of documents. The estimation of $\widehat{\theta}_d$ is obtained from the feature frequencies encoded by each document vector. Finally, the integration of the kernel function in the binary SVM classifier allows to define decision boundaries which maximize the margins derived from the geodesic

[5] Throughout this chapter, n+1 denotes the dimension of the ambient space where is immersed the features space. This dimension is given by the cardinal of the vocabulary of the VSM model.

distances.

As can be seen, the methods *Multinomial Manifold* proposed in [Lebanon, 2005] and [Zhang et al., 2005] are using the intrinsic geometry of the distributions space, instead of the geometry of the data on the features space. Curiously, in the features space we have random vectors for each class, and these same vectors have been associated to random vectors of distributions, but without to consider that the classes are generated by a probability distribution. They have transformed a classification problem of features vector into one about the classification of distributions, it means, they have converted the parameters of the distributions in a new set of features.

From a geometric point of view, a thoughtful reading of the ideas above turns very interesting. We appreciate some way of duality among the features space and the distributions spaces that is not trivial at first glance. In our method, we represent the classes by geometric distributions on the features space, the ideas above suggest that one interesting variant would be to represent the classes as a random vector on the multinomial distributions, following the ideas provided by Lebanon [Lebanon, 2005] and Zhang et al [Zhang et al., 2005]. If we try this approach, we can transform our intrinsic Bayes classifier on the features space in an intrinsic Bayes classifier on the multinomial distributions space, defining a future trend to be explored.

By analyzing the geodesic equations on the distributions space and the normalizations of the vectors in the *Multinomial Manifold* method, we appreciate a clear parallelism with some ideas expressed in our method, regardless of differences in space and representation of classes. Our intuition is that we should expect similar results.

Although the idea about the use of a kernel function with the SVM method in [Lebanon, 2005] and [Zhang et al., 2005] is not new [Hofmann et al., 2008], the approach is good because it allows a smooth integration of novel methods with well accepted methods and libraries.

Using the kernel's theory [Hofmann et al., 2008], it is possible to prove that the *geodesic Mahalanobis distance*, introduced in this thesis, also defines a strictly positive kernel. In this way, this result links with the kernels, what allows to use the theorems and results from functional analysis about the subject. The use of the *geodesic Mahalanobis distance* as distance measure among vectors on the features space implies to assume that all the a priori probabilities of the classes are equal, thus, in this case, we would be using only the geometry of the distribution.

3.2.2.2 Geometry of the features space: kNN-data-manifolds

The methods in this section are characterized by trying to represent the geometry of the classes on the features space, and they have been called *data-manifolds* methods. By other hand, most of them try to integrate in their models some invariant local property, whose aim is to preserve the geometric relations in the local neighborhood of the data. To capture these local properties, the methods use typically some sort of kNN-based neighborhood structure, therefore we have added the prefix kNN to the name of this family of methods. Some of the features share by the methods in this family are as follows:

1. The geometric structure of the data is represented by a neighborhood graph

of kNN type. The points are identified like adjacent ones according to its Euclidean distance on the features space.

2. The local structure is approximated by the initial kNN graph, while the final classification uses any variant of the kNN method with some sort of geodesic distance, measured on the adjacency graph, like distance for the classification. The core idea of the work in [Boczko & Young, 2005] is to measure the distance among sets is also similar to the kNN methods, but this method is far from the first ones in some aspects.

As example of the methods in the kNN-data-manifold family, we can cite: MKNN [Wen et al., 2006], MFA [Yan et al., 2007], LLE [He et al., 2008], KDGPP [Wang & Chen, 2009], MDA [Wang & Chen, 2009] and MAED [Cai & He, 2012].

Bockzo et al [Boczko & Young, 2005] [Boczko et al., 2009] introduce a general-purpose classification method, not used before in the TC problem, which tries to represent the geometry of the data on the ambient space. The proposed method represents the decision boundaries through a *signed distance function* (SDF), which can be interpolated by radial basis functions (RBF) [Buhmann, 2003] on the features space. The distance among the data points and the decision boundaries is approximated through a distance function among any point and its complementary set (positive/negative), thus, the decision boundaries are implicitly defined by the zero level set of the interpolant RBF function [Buhmann, 2003]. The classifier is of binary type, and the distance $d\left(x, \overline{C}\right)$ from a positive point $x \in C^{+}$, or negative point $x \in C^{-}$, is approximated by the half value of the minimum distance from the point to its complementary set $\overline{C}$, such as is shown in (3.3).

$$d\left(x, \overline{C}\right) = \min_{j} \left\{ \frac{1}{2} \left\| x - x_j \right\|, \forall x_j \in \overline{C} \right\} \tag{3.3}$$

The method of Bockzo et al. [Boczko et al., 2009] is defined on the Euclidean space, but it could be adapted to the TC problem whether the document vectors are represented on the unit hypersphere and the geodesic distance is used to approximate the signed distance function. Unlike our method, this method has not been applied before to the TC problem, therefore, it does not consider the structure of the features space in this case, and its geometric model for the data does not represent the data by a distribution on the features space, such as we do in our model. In [Boczko et al., 2009], the authors propose like future work to improve the computation of the distance among subsets and the geometric model for the classes. Precisely, the possibilities and research lines introduced in [Boczko & Young, 2005] [Boczko et al., 2009] us to start the research carried-out in this chapter.

In [Wen et al., 2006], Wen et al. introduce a TC method based in a variant of the kNN algorithm, which is called MKNN. The variant consists in the definition of a geodesic distance on the dataset, which tries to capture the intrinsic geometry of the whole set. For this, they build a distance matrix encoding the distance among each pair of documents (x, y), where the geodesic distance $d_g(x, y)$ is approximated by the shortest path with degree 2 (2 edges) over the set, measured by the Euclidean metric (chord in the ambient space), such that $\widehat{d_g}(x, y) = \min_{z} \{d_e(x, y), d_e(x, z) + d_e(z, y)\}$. The core idea of the method is to

capture the geometry of the dataset through the shortest path on the initial adjacency graph of the data. The proposed solution is simple, but limited. Formally, the approximation for the geodesic distance should analyze all the possible paths on the graph of the set, what would be intractable for large datasets, because it is equivalent to the computation of the Dijkstra algorithm among all pairs of points. Moreover, they do not consider the intrinsic geometry of the features space, but approaching the geodesic arcs by the distance in Euclidean ambient space. Their geometric representation either considers that the data could be generated by a distribution, therefore, the shape of the dataset is defined by a discrete set of random samples which does not match the intrinsic geometry of the distribution. The last problem is shared by all the kNN algorithms.

One important feature of the kNN classifiers is that they converge to the optimum Bayes classifier when the dataset is sufficiently dense, what occurs when every cell on the features space contains a sample set that matches the proportional distribution among classes. The MKNN method improves the results of an standard kNN method that uses the cosine function like similarity measure. As we will see in section 3.3, the cosine function matches the inverse of the geodesic distance on the hypersphere, and it induces the same ranking order on the data. The improvement introduced by MKNN suggests that any way of encoding of the geometry of the data in the models can improve the performance of the resulting classifier, although this improvement could look simple.

In [Yan et al., 2007], Yan et al. introduce a method for the dimensionality reduction of the features space in *image classification*, called *Marginal Fisher Analysis* (MFA), which is based in the definition of a representation of the geometry of the dataset for a same class. The method builds a structure called *Graph Embedding*, which represents the geometry of the data through kNN clusters of near points. The authors consider their method like a variant of the LDA method.

He et al. [He et al., 2008] introduce a dimensionality reduction method for the TC problem which is based in the *Linear Local Embedding* (LLE) method developed in the context of data mining [Roweis & Saul, 2000]. In the LLE model, every point is represented as a barycentric combination of its k-neighbors. The k-nearest neighbors are defined using a metric on the Euclidean ambient space, then, the adjacency relations allow to build a graph that tries to approximate the shape of the data. The idea behind LLE is to approximate the local neighborhood of every data point by a flat Euclidean subspace with lower dimension while the adjacency relationships are preserved in the deformed representation. LLE tries to represent the geometry of the data using local approximations without to make any assumption about the data distribution, thus, in this sense, LLE is a local manifold method. Finally, the classification is carried-out over the transformed data obtained through LLE, through a kNN algorithm that uses the Euclidean distance on the lower dimension space as metric for the classification. In [He et al., 2008], the authors make a comparative analysis between LLE and LSI, and they admit that the dimensionality reduction produced by LSI (derived from a SVD decomposition) is higher that LLE, although LLE allows a better representation of the geometry of the data. This method does not consider the geometry of the features space where the data are embedded, and it does not represent the geometry of the data by an intrinsic distribution like we propose in our method. By last, the method does not consider that the data are

already embedded in a differential manifold (the features space) that is a subset of the Euclidean ambient space.

In [Cai & He, 2012], Cai and He introduce an active learning method for TC called MAED, which is based in the representation of the geometric structure of the data. The active learning methods (1) have as main goal the selection of the best samples for the training of any classifier, and in most of cases, they are independent of the chosen classifier. In statistics, the sampling problem is known like *optimal experimental design*. The local geometry of the data is captured through a neighborhood graph that is computed using the Euclidean distance in the features space. Later, the method builds a kernel function based in a Laplacian matrix of the graph [6], which represents the distance relations and adjacency among the data. By last, the authors evaluate experimentally their ideas with some variants of SVM method, proving that the samples selection improves the classifier results.

Problem 1 (Active learning) *Given a dataset $\mathcal{X}$ represented by vectors in an Euclidean space $\mathcal{X} = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^m$, find the $\mathcal{Z} = \{z_1, z_2, \ldots, z_k\} \subset \mathcal{X}$ with the most informative elements.*

In [Wang & Chen, 2009], Wang and Chen introduce a method for *image classification* called *Manifold Discrimant Analysis* (MDA). The method is similar to other methods like LLE [He et al., 2008] and LDA [Torkkola, 2004], and it also computes distance among local datasets. The method transforms the local neighborhoods to a lower dimension space, then, it defines the decision boundaries that maximize the margin among classes. This approach is similar to the distance function proposed by Boczko et al. [Boczko & Young, 2005], and to the margin concept in SVM [Cortes & Vapnik, 1995]. The MDA method builds a class hierarchy using a kNN-type graph based that combines Euclidean distances d_e and geodesic distances d_g, such as the graph proposed by MFA [Yan et al., 2007] , and how is made by most of the manifold-based methods. Once the Euclidean and geodesic distance matrixes for the data have been built, a novel matrix is computed whose entries are the factors $\frac{d_g}{d_e}$. The aim of this novel matrix is to measure the non-linearity (curvature) of the dataset in the features space. The non-linearity measure is used to split the dataset in clusters of uniform curvature. Using image processing terminology, the method segmentates the images using an approximation for the curvature of the geometric set derived from the data values. Each image class is represented by a collection of clusters resulting from the segmentation, then, the distance among different images is defined by the pairwise distance among the clusters of each one.

In [Wang et al., 2012], Wang et al. introduce a novel extension to a dimensionality reduction method, called DGPP [Song & Tao, 2009], which is proposed in image processing. The proposed method is called KDGPP, and it is based in the representation of the intra-class geometry and the discriminant information among different classes. DGPP uses a dimensionality reduction technique similar to LLE [He et al., 2008]. Like the LLE method, DGPP also builds a weighted adjacency graph to represent the local invariance, but surprisingly, no work cites the other

[6] The Laplacian matrix of a graph is a well known structure in the scope of the geometric modeling, where it is used to represent deformation and smoothing functions on meshes. This matrix is equivalent to a diffusion process (filtering) driven by a Laplace equation in partial derivatives.

one, maybe because they have been developed in different fields. The main novelty introduced in [Wang & Chen, 2009] is the reformulation of the DGPP method through a radial basis kernel function (RBF) to estimate the weights for the adjacency graph, using the Euclidean distance among vectors like proximity measure to build the kNN neighborhoods. Once the data are embedded in the lower dimension space, the classification is made using the kNN algorithm and the Euclidean distance among the transformed data, such as is made by the LLE-based method proposed in [He et al., 2008].

The main drawback of the models described in this section is that they use non-smooth discrete approximations to represent the local geometry of the classes. Although these approximations produce some improvements in the precision of the classifiers, they restrict the generalization of the classifiers to unseen samples, or regions not densely sampled. Moreover, they share some known problem with the classical kNN classifiers, such as the sampling uniformity and density, and the lack of smoothing. Curiously, these methods build the adjacency graphs using the Euclidean distance in the ambient space, instead of the intrinsic distance of the true features space.

Our method represents the geometry of the classes like an intrinsic distribution on the geometry of the features space, where it is defined by the positive unit hypersphere. In our method, the geometry of the dataset of a class is represented by a smooth generative model given by the multivariate Gaussian distribution. This approximation is equivalent to represent the geometry (shape) of the dataset by a quadric hypersurface embedded in the features space, what counteracts the overfitting effect of the discrete models. The Gaussian distribution acts like a smoothing filter on the geometry of the dataset. Our representation with intrinsic Gauss functions induces a smooth model, continuous and differentiable on the features space, which allows a better generalization capability, in special in regions with low sampling density. Moreover, the use of intrinsic Gaussian functions is endorsed by the Central Limit's theorem, because the Gaussian mixture is the limit for the kNN classifiers when the training set is dense, precisely, this property suggests to use this limit distribution instead of a discrete version to approximate the geometry of the data.

On the other hand, our hypothesis respect to the representation of the data by an intrinsic normal mixture also captures the geometry of the features space, what has not been considered before by other methods. Conversely, the rest of manifold-based methods make use of the Euclidean distance in the ambient space to build the adjacency graphs that later are used to approximate the local geometry of the classes.

3.3 Preliminaries

In this section, we make a detailed review of some aspects of the vector space model that will be of interest in the development of work. Moreover, we also introduce the geometry of the features space, including the notation and definitions needed to follow the exposition of our model.

Throughout our exposition, we always use lowercase letters to denote vectors,

and uppercase letters for sets. The notation a^T denotes the transpose of any vector a, while that the notation $a^T b$, or $a^T \cdot b$ denotes the usual scalar product among vectors.

3.3.1 Vector Space Model (VSM)

As first step for the classification of text documents, the documents are represented using the known vector space model (VSM). A vector space model is defined by the tuple $VSM = (D_T, V, \lambda)$ as follows: (1) D denotes the documents space and D_T the set of documents used for the training of the model; (2) V denotes the vocabulary of keywords, which can be composed by stemmed words obtained by any stemming method like the proposed one in [Porter, 1980], or features with a higher lexical-semantics meaning [Chua & Kulathuramaiyer, 2004] [Khan et al., 2010], such as phrases, root words, semantic classes and synonyms obtained from ontologies and thesaurus; and (3) a mapping function λ which embeds the documents in the vector space.

$$\lambda : D \rightarrow M \subset \mathbb{R}^{n+1} \tag{3.4}$$

The function λ maps every document $d \in D$ to a vector x defined on a subset of the vector space $\mathbb{R}^{n+1}$, what we call *features space* and denote by the letter M. The mapping λ is not injective, because two documents can be mapped to the same vector. The features space M is embedded in the *Euclidean ambient space* $\mathbb{R}^{n+1}$, whose dimension is defined by the cardinal of the vocabulary V, being $n+1 = |V|$.

Every document $d_i \in D$ is transformed by λ in a vector $x_i = \left(x_i^1, \ldots, x_i^{n+1}\right) \in M$, whose components x_i^j are called weights. The weights are computed through a weighting function of the occurrence frequency of the elements in V within the document d_i and the training corpus D_T. The most popular weighting function is the TFIDF, although there are many other variants, such as the described ones in the thesis of Fresno [Fresno, 2006].

The document vectors have some properties that we summarize as follows:

1. All the components x_i^j are positive real values, because these values correspond to the frequency values of the vocabulary terms, therefore $x_i^j \geq 0, \forall j \in J$, where J is the coordinates index set $J = \{1, \ldots, n+1\}$ for each dimension of the ambient space.

2. The vectors are normalized to make them invariant with regard to the document length, such that $\|x_i\| = 1, \forall d_i \in D$. As consequence of the normalization, the features space is reduced to the positive part of the unit hypersphere in $\mathbb{R}^{n+1}$, denoted by S^n. Specifically, the features space M is defined by $S_+^n = \left\{x \in \mathbb{R}^{n+1} : x^T \cdot x = 1, x^j \geq 0, \forall j \in J\right\}$.

3. The structure of the vectors is sparse what means that most of the coordinates are zero. This is a consequence of the non-occurrence of every vocabulary term within a document, which follows from the use of features selection method to increase the discriminant capability among classes. By example, the vectors in the corpus RCV2 [Lewis et al., 2004] have a dimension of 47236, however, the average number of non-zero coordinates in the training set is 61 and the maximum 164.

4. The vectors of a same class are randomly distributed on the features space S_+^n. The selection of discriminant terms for the global vocabulary and the sparse structure of the vectors contributes to the observation of *outliers*.

5. The number of features $|V|$ of the VSM model is proportional to the number of classes. It follows from the necessity to include features to separate different classes and to keep close the documents in a same class, what requires to increase the vocabulary whenever the number of classes is increased.

One of the main problems of the keyword-based VSM models, like the studied in this chapter, is the difficulty to build vocabularies with wide coverage of the corpus, and good generalization and discriminant capabilities. The definition of the vocabularies for the VSM models is a problem of features selection, and it is the main source for the sparsity and the null similarity among most of vector pairs.

Other known drawback of the keyword-based VSM models is the lack of meaning for the terms in its vocabulary, which decreases its generalization capability. This problem manifests in the difficulty to retrieve documents using terms not mentioned in the vocabulary or the corpus.

Precisely, these problems are solved by the recent family of ontology-based IR models, such as the model called Intrinsic Ontology Spaces that we introduce in the chapter 2.

Most of TC methods are based in supervised machine learning algorithms. For this, the training corpus D is divided in two subsets $D = D_E + D_T$. The set D_E for training and the set D_T for tests.

3.3.2 Geometry of the features space

The features space M in the VSM model used for text categorization is defined by the positive part of the unit hypersphere S^n, being denoted by S_+^n in (3.5).

$$S_+^n = \left\{x \in \mathbb{R}^n : \ x^T x = 1, x_i \geq 0\right\} \tag{3.5}$$

Throughout the chapter, we use the concept of differential manifold for the features space S_+^n. For sake of completeness, we have included here some basic definitions that the reader could also find in some introductory texts as [do Carmo, 1992] and [Gamboa & Ruiz, 2006].

Definition 22 (Differential manifold) *A differential manifold of dimension n is a set M and a family of injective mappings $\{x_\alpha : U_\alpha \subset \mathbb{R}^n \to M\}$ of open subsets U_α of $\mathbb{R}^n$ into M, such that:*

(1) *Covering*: $\cup_\alpha x_\alpha (U_\alpha) = M$

(2) *Differentiable change of coordinates*: for any pair α, β, with $x_\alpha (U_\alpha) \cap x_\beta (U_\beta) = W \neq \emptyset$, the sets $x_\alpha^{-1}(W)$ and $x_\beta^{-1}(W)$ are open sets in $\mathbb{R}^n$, and the mappings $x_\beta^{-1} \circ x_\alpha$ are differentiable.

(3) *Atlas maximal*[7]: the family of sets $\{(U_\alpha, x_\alpha)\}$ is maximal with regard to the conditions (1) and (2).

A *differential manifold* is a structure that defines a set of local parameterizations x_α over a set M, called *charts*, with the property that the changes of coordinates among overlapping charts are *differentiable*. If the differentiability degree is not said, it will be assumed of class C^∞. The family of charts X is called *atlas* and when it verifies the conditions (1) and (2) is called *differentiable structure*. The *differential manifolds* are locally diffeomorphic spaces [8] to subsets of the affine space $\mathbb{R}^n$, which allow to apply the notions of differential and integral calculus on the underlying set M. The dimension of a manifold is defined by the dimension of its charts.

3.3.2.1 Unit hypersphere parametrization

The features space S_+^n can be parameterized using a unique chart $\mathcal{X}$ in polar coordinates (3.6), by removing the pole $(1, \ldots, 0_{n+1}) \in \mathbb{R}^{n+1}$. The removal of the pole guarantees that the chart is an injective mapping and it verifies the axioms described above. In figure 3.1, we show a diagram of the construction.

$$\begin{aligned} \mathcal{X} \quad &: \quad \Omega \to S_+^n \\ \mathcal{X}(u) \quad &\longmapsto \quad \begin{bmatrix} x_1(u) \\ x_2(u) \\ \vdots \\ x_{n+1}(u) \end{bmatrix}_{(n+1)\times 1} \\ u \quad &= \quad (u_1, \ldots, u_n) \in \Omega \end{aligned} \tag{3.6}$$

The domain of S_+^n is defined by $\Omega = (0,1)^n$, while that the formula (3.7) defines the coordinate functions $x_i(u)$.

$$\begin{aligned} x_1(u) &= cos\left(\frac{\pi}{2}u_1\right) \\ x_2(u) &= sen\left(\frac{\pi}{2}u_1\right) cos\left(\frac{\pi}{2}u_2\right) \\ x_3(u) &= sen\left(\frac{\pi}{2}u_1\right) sen\left(\frac{\pi}{2}u_2\right) cos\left(\frac{\pi}{2}u_3\right) \\ &\vdots \\ x_n(u) &= sen\left(\frac{\pi}{2}u_1\right) \cdots sen\left(\frac{\pi}{2}u_{n-1}\right) cos\left(\frac{\pi}{2}u_n\right) \\ x_{n+1}(u) &= sen\left(\frac{\pi}{2}u_1\right) \cdots sen\left(\frac{\pi}{2}u_{n-1}\right) sen\left(\frac{\pi}{2}u_n\right) \end{aligned} \tag{3.7}$$

[7] A maximal atlas is an equivalence class for all the possible combinations of charts that cover the same base space, formalizing the fact that a manifold is the same with regard to any valid parameterization.

[8] A diffeomorphism is a bijective mapping among topological spaces such that the direct mapping and its inverse are continous and differentiable.

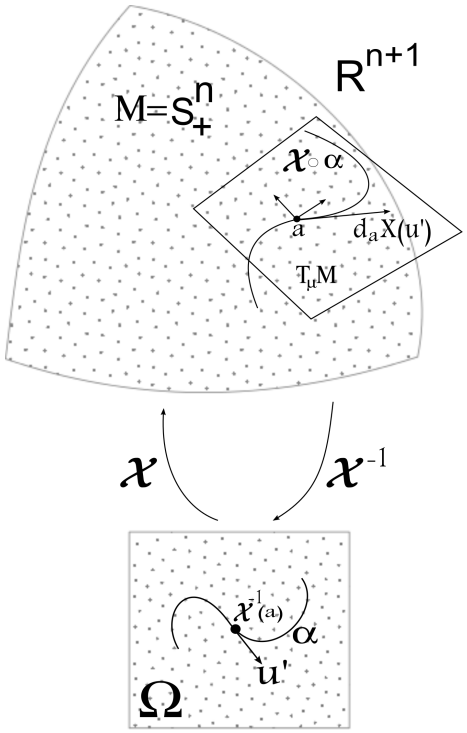


Figure 3.1: Parametrization $\mathcal{X}$, tangent space T_aM and differential map $d_a\mathcal{X}$ for the unit hypersphere S_+^n.

The inverse parameterization $\mathcal{X}^{-1} : S_+^n \to \Omega$ is defined by (3.8).

$$\begin{aligned} u_1 &= arccot\frac{x_1}{\sqrt{x_2^2+\cdots+x_n^2+x_{n+1}^2}} \\ u_2 &= arccot\frac{x_2}{\sqrt{x_3^2+\cdots+x_n^2+x_{n+1}^2}} \\ &\vdots \\ u_{n-1} &= arccot\frac{x_{n-1}}{\sqrt{x_n^2+x_{n+1}^2}} \\ u_n &= 2\cdot arccot\frac{x_n+\sqrt{x_n^2+x_{n+1}^2}}{x_{n+1}} \end{aligned} \tag{3.8}$$

3.3.2.2 Geodesics on the hypersphere

The geodesics on the unit hypersphere match the maximal circles, therefore the geodesic distance d_g is the arc-length of the shortest geodesic arc among two points, whose value is given by (3.9). Note that this formula is the same for a circle (S^1), a sphere (S^2), or any n-sphere (S^n).

$$\begin{aligned} d_g &: S^n \times S^n \to \mathbb{R}^+ \cup \{0\} \\ d_g(a,b) &= \arccos\left(a^T \cdot b\right) \end{aligned} \tag{3.9}$$

3.3.2.3 Tangent space of the features space

Being M a differential manifold, its tangent space in a point $a \in M$ is denoted by T_aM, being an affine space with dimension n. The tangent space T_aM is tangent to the manifold on the base point $a \in M$, such as is shown in figure 3.1.

The tangent space in a point $a \in M$ is generated by a set of base vectors $\left\{\frac{\partial \mathcal{X}}{\partial u_i}\right\}$, which are defined by the images of the vectors in the canonical base of $\mathbb{R}^n$ through the linear mapping $d_a\mathcal{X}$ (3.10). The mapping $d_a\mathcal{X}$ is called the differential of the manifold. The canonical base vectors are $e_i = (0, \ldots, 1, \ldots, 0)$, where each 1 is defined in the i-th position. The linear mapping $d_a\mathcal{X}$ (3.10) is defined by the Jacobian (3.12) of the parametrization $\mathcal{X}$.

$$\begin{array}{rcl} d_a\mathcal{X} & : & \mathbb{R}^n \to T_aM \subset \mathbb{R}^{n+1} \\ d_a\mathcal{X}(u') & \longmapsto & J \cdot u' \end{array} \tag{3.10}$$

The set of vectors $\left\{\frac{\partial \mathcal{X}}{\partial u_i}\right\}$ define a base for the tangent space T_aM, whose definition is given by (3.11)

$$\frac{\partial \mathcal{X}}{\partial u_i} = \begin{bmatrix} \frac{\partial x_1}{\partial u_i} \\ \vdots \\ \frac{\partial x_{n+1}}{\partial u_i} \end{bmatrix}_{(n+1)\times 1} \tag{3.11}$$

The differential $d_a\mathcal{X}$ defines a parametrization of the tangent space T_aM, mapping the tangent directions in the domain, denoted by $u' \in \mathbb{R}^n$, to the tangent space T_aM. The vector u' is interpreted like the tangent to any curve $\alpha(t) \in \Omega$ in the point $\mathcal{X}^{-1}(a)$, such as is shown in figure 3.1.

$$J = \begin{bmatrix} \frac{\partial \mathcal{X}}{\partial u_1} & \cdots & \frac{\partial \mathcal{X}}{\partial u_n} \end{bmatrix}_{(n+1)\times n} \tag{3.12}$$

$$J = \begin{bmatrix} \frac{\partial x_1}{\partial u_i} & \cdots & \frac{\partial x_1}{\partial u_i} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_{n+1}}{\partial u_i} & \cdots & \frac{\partial x_{n+1}}{\partial u_i} \end{bmatrix}_{(n+1)\times n}$$

3.3.3 Bayes classifier

Given a set of document classes $C = \{c_k\}_{k\in K}$ we can represent every class c_k by an intrinsic normal distribution $\mathcal{N}_g(\mu, \Sigma_g)$ on the features space, whose d.p.f. is the function $f_{g,k}(x : \mu_k, \Sigma_{g,k})$ in (3.13), defining the conditional probability $p(x \mid c_k)$, whose meaning is the observation probability for $x \in S_+^n$ when it has been generated by the class c_k (hypothesis).

$$\begin{array}{rcl} p(x \mid c_k) & = & f_{g,k}(x, \mu_k, \Sigma_{g,k}) \\ p(x \mid c_k) & = & (2\pi)^{-\frac{n+1}{2}} |\Sigma_{g,k}|^{-\frac{1}{2}} e^{-\frac{1}{2}r_{g,k}^2} \end{array} \tag{3.13}$$

The total probability that the whole set of classes C could generate a vector $x \in S_+^n$ is given by the formula (3.14), where $p(c_k)$ denotes the a priori probability for the class c_k.

$$p(x) = \sum_{k\in K} p(x \mid c_k)\, p(c_k) \tag{3.14}$$

Given an observed vector $x \in S_+^n$, we can use the Bayes's theorem to compute the a posteriori probability $p(c_k \mid x)$ that a given class c_k would have generated it, such as is given by (3.15).

$$p(c_k \mid x) = \frac{p(x \mid c_k) \cdot p(c_k)}{p(x)} \tag{3.15}$$

The function $p(c_k \mid x)$ defines a ranking value, or "score" for the observation x, and according to the theory of Bayes, the decision that maximizes the probability to observe x is the selection of the class that maximizes $p(c_k \mid x)$, also know as maximum a posteriori (MAP). It leads us to the optimal Bayesian classifier in (3.16).

$$c^* = \underset{c_k \in C}{\arg max} \{p(c_k \mid x)\} \tag{3.16}$$

As we will see in next section, the proposed method defines a Bayesian classifier using a normal distribution to represent the data constrained by the geometry of the features space. In this sense, the proposed method is an optimal Bayes's classifier on a differential manifold.

3.4 Intrinsic Bayes-Voronoi classifier

In this section we introduce a document classification method that is called *Intrinsic Bayes-Voronoi classifier* (IntBayesVor). The chosen name refers the fact that the method selects the optimal Bayes solution for a probability distribution defined on the features space, which induces an intrinsic Voronoi diagram. The definition of the model integrates the intrinsic geometry of the TC problem as is defined by the VSM model, namely: (1) the geometry of the features space given by the positive unit hypersphere, and (2) the geometric representation, or approximation, for the classes.

The geometry of every class is defined by a geodesic normal distribution $\mathcal{N}_g(\mu, \Sigma_g)$, which expresses the geometry of the regions of a class using the geodesic distance on the feature space, resulting in a function that we call *geodesic Mahalanobis's distance* r_g.

The concepts $\mathcal{N}_g$ and r_g are derived in section 3.4.1 as a generalization of the normal distribution on a differential manifold. Although we are not aware of the definition of these objects in statistics, is quite likely to have been previously defined, so that it is necessary a deeper review of the literature to trace their origins, applications and properties. In any case, although these concepts would have been previously defined, they have never been used to model the intrinsic geometry of the TC problem and to obtain the optimal classifier introduced in this work, so it does not detract merit the novelty presented here.

Starting from the geodesic normal distribution $\mathcal{N}_g$, we define the geodesic log-linear likelihood function $\mathcal{L}_g(x : \mu, \Sigma_g)$, function that allows us to define the local distance[9] denoted by $\mathcal{B}_g^k(x : \mu, \Sigma_g)$, that we call *Bayes's distance*. The Bayes's

[9] As we will see in its definition, this is not a traditional metric that applies for every pair of elements in the space, but for each element of the space respect to a set of distinguished elements.

distance defines the probabilistic distance among every point on the features space and the centroid μ_k of each class c_k.

We introduce some theorems to prove that the function $\mathcal{B}_g^k(x:\mu,\Sigma)$ defines the optimal solution for the intrinsic classification according to the Bayes's criterion, and it induces a Voronoi diagram on the features space, denoted by $\mathcal{V}_C$.

One theorem proves that the intrinsic Bayes solution always induces a Voronoi diagram on the features space as differential manifold, whose boundaries are defined by the intrinsic geometry of the features space, the data distribution and the a priori probabilities of the classes.

We also note that the definition of $\mathcal{V}_C$ applies to any features space M which is defined as a differential manifold, although here we only focus in the specific case for the hypersphere S_+^n with an intrinsic normal distribution. The use of the hypersphere S_+^n and the normal distribution $\mathcal{N}_g$ leads us to get closed formulas for all the involved objects.

The intrinsic Voronoi diagram $\mathcal{V}_C$ is expressed implicitly by the set of functions $\mathcal{B}_g^k\left(x:\mu_k,\Sigma_g^k\right)$, which means that we do not need explicitly calculating its representation what would imply computing and storing an explicit representation of the decision boundary, being an intractable [Dwyer, 1989] [Boissonnat & Karavelas, 2002] and unnecessary problem in our case.

Once the structure of our classifier is defined, we present in the section 3.4.5 the numerical methods and algorithms needed to compute it. The training process model consists in the estimation of the a priori probability $p(c_k)$ for each class c_k and its intrinsic distributions $\mathcal{N}_g$, what allows to define the functions $\mathcal{B}_g^k\left(x:\mu_k,\Sigma_g^k\right)$.

For the training of the classifier we introduce a method and algorithms which could be improved in the future, but they will continue using the main result of this chapter: the fact that the solution to the TC problem can be defined as an optimal Bayes classifier on the geometry of the features space, under the hypothesis that the classes match the geodesic normal distribution.

Finally, the proposed method is a simple Bayes's classifier, such that for a novel unseen document vector $x \in S_+^n$, the selection of the optimal class $c^* \in C$ is only a search process to find the argument that minimizes the functions $\left\{\mathcal{B}_g^k\left(x:\mu,\Sigma_g^k\right)\right\}_{k\in K}$.

The remainder of the section is structured as follows. In section 3.4.1, we introduce a detailed description of the representation model for the geometry of the classes. The section 3.4.2 introduces the definition of the intrinsic Bayes's distance. In section 3.4.3, we show how to classify an unseen document. The section 3.4.4 introduces the our definition for the intrinsic Voronoi diagram and its relation with the optimal Bayes solution. By last, in section 3.4.5 we propose some algorithms to implement the proposed classifier.

3.4.1 Representation of the classes

In this section, we introduce a geometric model for the representation of the subsets (regions) of the features space $R_k \subset M = S_+^n$, where the data generated by a given class $c_k \in C$ are contained with some probability. We will assume that the vectors generated by every class can be represented by an intrinsic normal distribution $\mathcal{N}_g^k(\mu_k,\Sigma_k)$ on the features space manifold M. The figure 3.2 shows some samples

distribution for a group of classes on the unit hypersphere (features space).

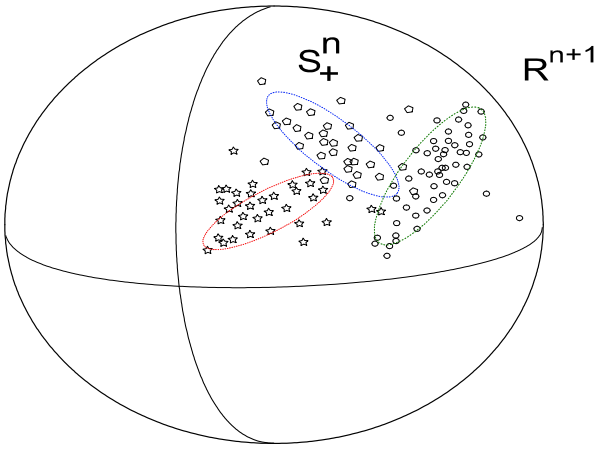


Figure 3.2: Clases sobre el espacio de rasgos S_+^n

The normal distribution for a random vector $x \in \mathbb{R}^{n+1}$ is given by (3.17), where $\mu \in \mathbb{R}^{n+1}$ is the mean vector and $\Sigma \in \mathbb{R}^{n+1 \times n+1}$ the covariance matrix for the normal distribution associated to a class.

$$\begin{aligned} f(x, \mu, \Sigma) &= (2\pi)^{-\frac{n+1}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \\ p(x \mid c) &= f(x, \mu, \Sigma) \end{aligned} \tag{3.17}$$

The function $f(x, \mu, \Sigma)$ represents the p.d.f.[10] for a random vector x on the Euclidean space $\mathbb{R}^{n+1}$, corresponding to a normal distribution denoted by $\mathcal{N}(\mu, \Sigma)$. This function defines the conditional probability $p(x \mid c)$ to get the vector x generated by class $c \in C$. The term in the exponent place in (3.17) is a quadratic form that allows to define a space of n-dimensional quadrics parametrized by (μ, Σ), and described by (3.18). The r value in (3.18) is the well known Mahalanobis distance. Because all the components of any VSM document vector x are positive, the covariance matrix Σ is symmetric, positive defined, and real-valued, thus, its inverse always exist. The last condition guarantees that it is always possible to fit the normal distribution $\mathcal{N}(\mu, \Sigma)$ to the vectors generated by any class in a VSM model, whenever an suitable dataset be available for its estimation.

$$(x - \mu)^T \Sigma^{-1} (x - \mu) = r^2 \tag{3.18}$$

The formula (3.18) defines level set surfaces for each r value, namely, the subset of the features space with the same occurrence probability value for the distribution of a given class.

The quadrics defined by (3.18) are sized and oriented according to the distribution of the data, property that we use to represent the geometry of the data on the features space.

[10] Probability density function.

Rewriting (3.17) as a function of the Mahalanobis distance, we get the simpler formula (3.19).

$$f\left(x:\mu,\Sigma\right)=\left(2\pi\right)^{-\frac{n+1}{2}}\left|\Sigma\right|^{-\frac{1}{2}}e^{-\frac{1}{2}r^{2}} \tag{3.19}$$

Up to now, we are only using the standard multivariate normal distribution. Now, we make a simple but key observation: in the VSM model used in the scope of the TC problem, the vectors are contained in the features space S_{+}^{n}, not in the whole Euclidean ambient space.

Starting from the observation above, we can define a novel intrinsic normal distribution that is denoted by f_g. The function f_g will be directly defined on the hypersphere, instead of the Euclidean ambient space $\mathbb{R}^{n+1}$ such as is defined in (3.19). The novel function f_g will represent the geometric distribution of the normal model associated to each class, through the use of the tangent space $T_{\mu}M$ to the features space S_{+}^{n}. The representation of the geometric distribution for the vectors generated by a class will be defined using the geodesic distance among points on the features space, and it will be defined on the tangent space of the hypersphere. The definition of the function f_g allows to join the intrinsic geometric of the data and the features space in a same mathematical object.

Now, observe the structure of the covariance matrix Σ for the Euclidean case. Every cell of the covariance matrix is defined by (3.20), where the factor σ is the joint standard deviation among the components of the vectors, and its square is the variance.

$$\begin{aligned}\Sigma^{ij} &= cov\left(x^{i},x^{j}\right)=\sigma_{ij}^{2}\\ &= E\left[\left(x^{i}-\mu^{i}\right)\left(x^{j}-\mu^{j}\right)\right]\end{aligned} \tag{3.20}$$

If we look closely at the term on right in (3.20), we find out other key fact in this study. The factors $(x^i-\mu^i)$ are encoding the expected value for the product of distances on each *tangent direction* in the base space (domain) of the p.d.f., in this case, the Euclidean space. Every factor $(x^i-\mu^i)$ measures the deviation in the direction of every canonical coordinate axis e^i.

Now, we are going to generalize the expression (3.20) to get a distribution whose domain is a *differential manifold* M, what for VSM reduces to S_{+}^{n}.

Given a vector $x\in M$, and being the features space M a differential manifold with parametrization X, the geodesic distance $d_g\left(x,\mu\right)$ is the measured distance on the manifold according to a geodesic outgoing from the point μ and passing through the point x, such as is shown in figure 3.3.

Fortunately, the geodesic on the unit hypersphere match the maximal circles, and we can compute its direction and length through a closed formula. We define the unit vector $a\in\mathbb{R}^{n+1}$ (3.21) on the chord that joins the points $\{\mu,x\}$.

$$a=\frac{1}{\left\|x-\mu\right\|}\left(x-\mu\right) \tag{3.21}$$

Then, we define the function g_{μ} (3.22) which assigns a vector in the tangent space $T_{\mu}M$ on any point $x\in X$ by the expression in (3.23). The vector $g_{\mu}\left(x\right)$ has

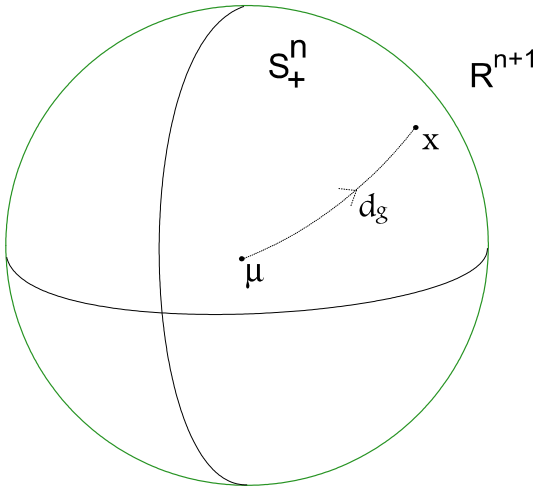


Figure 3.3: Distancia geodésica entre puntos

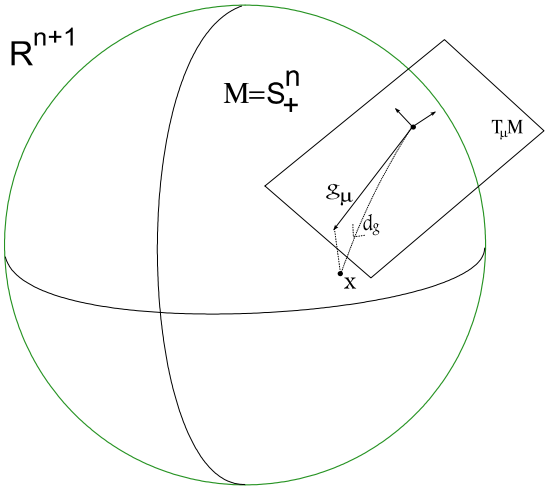


Figure 3.4: Representación del vector $g_\mu(x)$

the same direction than the outgoing geodesic from μ to x, and its Euclidean norm is $d_g(\mu, x)$. The definition of the function g_μ depends on the tangent space $T_\mu M$ and the family of outgoing geodesics at the point μ, such as is shown in figure 3.4.

The vector $g_\mu(x)$ is defined by (3.23), whose meaning is simply the projection of the vector (3.21) on the tangent space $T_\mu M$, taking advantage of that the unit normal in μ is equal to μ.

$$g_\mu : X \to T_\mu M \tag{3.22}$$

$$g_\mu(x) = \frac{d_g(\mu, x)}{\|a - (a^T\mu)\mu\|}\left(a - \left(a^T\mu\right)\mu\right) \tag{3.23}$$

The method of construction of the formula (3.23) could be generalized for any differential manifold M, however it is only valid for the hypersphere, because we are using the definition of the geodesic chord given by the vector $a \in \mathbb{R}^{n+1}$. For a general

manifold, the geodesic distance $d_g(\mu, x)$ should be obtained by the integration of one vector differential equation defined on any hypersurface with arbitrary dimension, problem that is extremely difficult in some scenarios as follows: non-convex surfaces, high-dimensional manifolds, multiple-chart manifolds, and any general surface with dimension greater than 2. In our case, we are lucky because the hypersphere offers the closed formula (3.24) for d_g, and substituting this value in (3.23) we get the closed formula for $g_\mu(x)$ in (3.25).

$$d_g(\mu, x) = arccos\left(x^T\mu\right) \tag{3.24}$$

$$g_\mu(x) = \frac{arccos\left(x^T\mu\right)}{\|a - (a^T\mu)\,\mu\|}\left(a - \left(a^T\mu\right)\mu\right) \tag{3.25}$$

Then, to get the standard deviations of $g_\mu(x)$ along each canonical direction in the tangent space $T_\mu M$, we just need to project $g_\mu(x)$ on each canonical base vector $\frac{\partial \mathcal{X}}{\partial u_i}$ de $T_\mu M$ and replacing in (3.20) to get the novel geodesic covariance matrix Σ_g^{ij} given by (3.26).

$$\Sigma_g^{ij} = E\left[\left(g_\mu^T(x) \cdot \frac{1}{\left\|\frac{\partial \mathcal{X}}{\partial u_i}\right\|}\frac{\partial \mathcal{X}}{\partial u_i}\right)\left(g_\mu^T(x) \cdot \frac{1}{\left\|\frac{\partial \mathcal{X}}{\partial u_j}\right\|}\frac{\partial \mathcal{X}}{\partial u_j}\right)\right] \tag{3.26}$$

The matrix Σ_g defines the geodesic covariance in the tangent space $T_\mu M$, what in our case encodes the variation of size and orientation of the data, measured on the unit hypersphere S_+^n.

Now, we are ready to define the novel geodesic density probability function f_g in (3.27). This function defines the intrinsic normal distribution of the data on the features space that we call *geodesic normal distribution* and is denoted by $\mathcal{N}_g(\mu, \Sigma_g)$. Similarly, the value of the quadratic form r_g in (3.28) is called *geodesic Mahalanobis distance.*

$$f_g(x : \mu, \Sigma_g) = (2\pi)^{-\frac{n+1}{2}}\left|\Sigma_g\right|^{-\frac{1}{2}} e^{-\frac{1}{2}r_g^2} \tag{3.27}$$

$$r_g^2 = g_\mu(x)^T\left[\widehat{J} \cdot \Sigma_g^{-1} \cdot \widehat{J}^T\right] g_\mu(x) \tag{3.28}$$

$$\widehat{J}_i = \frac{1}{\left\|\frac{\partial \mathcal{X}}{\partial u_i}\right\|}\frac{\partial \mathcal{X}}{\partial u_i} \in \mathbb{R}^n \tag{3.29}$$

The consequence of the expressions above is that the isoprobability regions (level sets) become a (n-1)-dimensional quadric defined on the features space S_+^n. In figure 3.5, we show the isoprobability curves for S_+^2 which define concentric ellipses embedded in the surface.

3.4.2 Intrinsic Bayes distance

Once the expressions (3.27) and (3.28) have been defined for the p.d.f. of the geodesic normal distribution, we will define the intrinsic Bayes distance, which allows us to define the optimal classifier in the next section. The Bayes distance characterizes

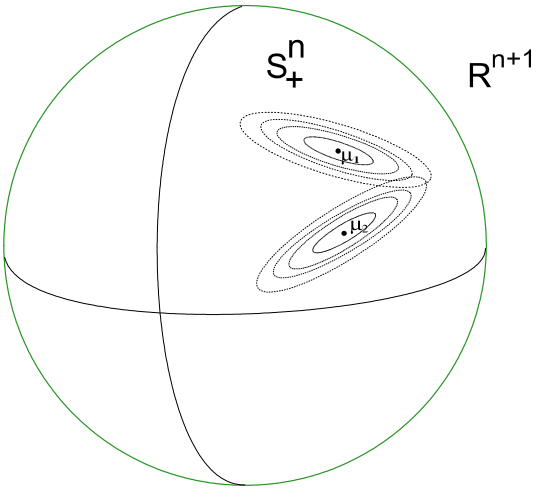


Figure 3.5: Regiones de iso-probabilidad de f_g

the geometry of the classification like a partition of the features space, thus finding a relationship between the optimal Bayes solution and the Voronoi diagrams on manifolds.

As we saw in section 3.3.3, according to the Bayes criterion, given any probabilistic distribution for the space of classes $C = \{c_k\}_{k \in K}$ and a document vector x, the optimal classification denoted by c^* is given by (3.30).

$$c^* = \underset{c_k \in C}{\arg max} \{p(c_k \mid x)\} \tag{3.30}$$

$$p(c_k \mid x) = \frac{p(x \mid c_k) \cdot p(c_k)}{p(x)} \tag{3.31}$$

For each class c_k we have the function $p(x \mid c_k) = f_g^k\left(x : \mu_k, \Sigma_g^k\right)$. Looking at the denominator in (3.31), we aware that $p(x)$ is constant for all the evaluations of $p(c_k \mid x)$, thus we can write the optimal solution as (3.32), what for the distribution $\mathcal{N}_g^k\left(\mu_k, \Sigma_g^k\right)$ reduces to the expression (3.33).

$$c^* = \underset{c_k \in C}{\arg max} \{p(x \mid c_k)\ p(c_k)\} \tag{3.32}$$

$$c^* = \underset{c_k \in C}{\arg max} \left\{p(c_k)\ f_g^k\left(x : \mu_k, \Sigma_g^k\right)\right\} \tag{3.33}$$

Finally, taking logarithms on both sides and doing some algebra on (3.27), we

get the likelihood geodesic log-linear function $\mathcal{L}_g$ (3.34), simplified in (3.35).

$$\begin{aligned} \mathcal{L}_g &: S_+^n \rightarrow (-\infty, -\delta] \subset \mathbb{R} \qquad (3.34) \\ \mathcal{L}_g(x) &= log\left(f_g^k\left(x : \mu_k, \Sigma_g^k\right)\right) \\ \mathcal{L}_g(x) &= -\frac{1}{2}\left[(n+1)\, log\,(2\pi)\right. \\ &\quad \left. + log\left(|\Sigma_g|\right) + r_g^2\right] \\ \mathcal{L}_g(x) &= -\delta - \frac{1}{2} r_g^2 \qquad (3.35) \\ \delta &= \frac{1}{2}\left[(n+1)\, log\,(2\pi)\right. \\ &\quad \left. + log\left(|\Sigma_g|\right)\right] \end{aligned}$$

Using the definition of $\mathcal{L}_g$, we can rewrite again the optimal solution by the expression (3.36).

$$c^* = \underset{c_k \in C}{argmax} \left\{ log\left(p\left(c_k\right)\right) + \mathcal{L}_g^k(x) \right\} \qquad (3.36)$$

As can be seen in (3.34),the function $\mathcal{L}_g(x)$ is monotonically increasing in the parameter x, but to define a distance function we need a monotonically decreasing function, thus, we will define the *intrinsic Bayes distance* for a class c_k, like the symmetric argument of (3.36), such as is shown in (3.37).

$$\begin{aligned} \mathcal{B}_g^k &: S_+^n \rightarrow \left[\delta - log\left(p\left(c_k\right)\right), \infty\right) \subset \mathbb{R} \\ \mathcal{B}_g^k(x) &= -log\left(p\left(c_k\right)\right) - \mathcal{L}_g^k(x) \qquad (3.37) \end{aligned}$$

Because $p\left(c_k\right)$ is a probability value, the next condition is verified: $p\left(c_k\right) \in [0,1] \rightarrow log\left(p\left(c_k\right)\right) \in (-\infty, 0]$.

By last, the optimal solution c^* is defined in (3.38) like a function of the novel intrinsic Bayes distance.

$$\boxed{c^* = \underset{c_k \in C}{argmin} \left\{ \mathcal{B}_g^k(x) \right\}} \qquad (3.38)$$

3.4.3 Classification of a document

The classification of an unseen document $p \in S_+^n$, reduces to the selection of the optimal class c^* according to the criterion (3.38), which we call *intrinsic Bayes rule*. We prove below that this solution always exists and its value is optimal and unique.

Theorem 2 (Intrinsic Bayes rule) *Be D a set of documents represented by points on a differential manifold M, and $C = \{c_k\}_{c_k \in C}$ a set of classes with geodesic normal distribution $\mathcal{N}_g\left(\mu_g, \Sigma_g\right)$ on M, and be $\mathcal{B}_g = \left\{\mathcal{B}_g^k(x)\right\}$ a family of Bayes distance functions associated to the classes, then the classifier $c^* = \underset{c_k \in C}{\arg\min} \left\{\mathcal{B}_g^k(x)\right\}$ always exists and its solution is the unique value that maximizes the posteriori probability $p\left(c_k \mid x\right)$.*

Proof:

1. *Existence.* The existence of the classifier is consequence of that the covariance matrix is real-valued, symmetric and positive defined, thus the inverse matrix Σ_g^{-1} always exists and it is unique.

2. *Uniqueness and optimal value.* Because $\mathcal{B}_g$ is a finite set, the set of real values $\left\{\mathcal{B}_g^k(x)\right\}$ is a finite subset of $\mathbb{R}$ with its usual order relation $(\mathbb{R}, \leq)$, thus, if exists a minimum, it must be unique. Moreover the value $\frac{e^{-\mathcal{B}_g^k(x)}}{p(x)}$ matches the maximum argument in the Bayes's theorem, thus the value c^* is optimal.□

The intrinsic optimal classifier (3.38) is optimal in the sense of Bayes, but it depends on the data fits to the distribution used as hypothesis, that in our case is the manifold-based Gaussian on the features space. Our hypothesis is endorsed by central limit theorem, despite that other distributions on the Euclidean space have been proposed in the literature. However, the fitting quality of the model to represent the data, as well as its estimation, will be always the main error sources, or deviations from the optimal solution.

3.4.4 Intrinsic Voronoi diagrams

The Voronoi diagrams are omnipresent structures in the scope of all sort of applications that require space partitions, such as: computational geometry, CAD, differential equations solving, etc. In [Aurenhammer, 1991], the author makes an excellent survey about the topic and its applications.

Next, we introduces an abstract definition of a Voronoi diagram that fits the context of our problem.

Definition 23 (Local distance) *Be X any non-empty set and an element $p_k \in X$ called site, and be $d_k : X \times p_k \rightarrow \mathbb{R}^+ \cup \{0\}$ a function defined on the neighborhoods B_{p_k} of p_k, such that $\forall b \in B_{p_k} \rightarrow b \subset X$ and $p_k \in b$. We said that d_k is a local distance function if it verifies the axioms below:*

1. Non-negativity: $0 \leq \epsilon_k \leq d_k(x, p_k)\,, \forall x \in X$

2. Coincidence: $d_k(x, p_k) = \epsilon_k \rightarrow x = p_k$

3. Symmetry: $d_k(x, p_k) = d_k(p_k, x)\,, \forall x \in X$

As we can see, the functions d_k are only defined respect to the site p_k and they are bounded below by a positive value ϵ_k. Moreover, the function does not support the triangle inequality, because it is only able to compute distances from any point to its site. Therefore, the functions d_k are not metrics, and they do not allows us to define families of open neighborhood [Arregui Fernández, 1988] to endow X with a structure of topological space, however, these functions are enough to define any Voronoi diagram as follows.

Definition 24 (Space with local distances) *Be X any non-empty set and $P = (p_k)_{k \in K}$ a set of elements in X called sites, and $S = (d_1, \ldots, d_k)_{k \in K}$ an ordered tuple of local distances associated to the elements in P by the indexes in K, then the tuple (X, P, S) is called a space with local distances.*

Definition 25 (Voronoi diagram) *Be the tuple (X, P, S) a space with local distances, then we call Voronoi diagram on X to every set $\mathcal{V} = (R_k)_{k\in K}$ that verifies the next axioms:*

1. $\mathcal{V}$ is a finite covering of the space X, induced by the set of sites P, such that $X = \underset{k}{\cup} R_k$ and $p_k \in R_k$, $\forall k \in K$.
2. Each set $R_k \subset X$ is called Voronoi cell and it is defined by the set of points in X whose distance to p_k is less than the distance to any other cell in $\mathcal{V}$, such as is expressed by (3.39).

$$R_k = \{x \in X : d_k(x, p_k) \leq d_j(x, p_j), \forall j \neq k\} \tag{3.39}$$

3.4.4.1 Intrinsic Bayes solution and the Voronoi diagrams

In this section we prove that the Bayes distance defined by (3.37) induces a Voronoi diagram $\mathcal{V}_C$ on the features space, and as consequence, the selection of the optimal class c^* according to the Bayes criterion reduces to find the Voronoi cell for the object to be classified.

Moreover, because the definition of the Bayes distance is independent of the probability function of the data, it is proven that any Bayes's classifier induces a Voronoi diagram on the features space, and the decision boundaries are the boundaries of the Voronoi diagram. Given a training dataset D_T defined according to the VSM model in section 3.3.1, we ask the following question:

> ¿ What is the geometric structure for the partition of a document set induced by the classifier in (3.38) ?.

Without more information, our intuition is that the best classification of the data should be one that selects as decision criterion the equidistant boundaries among classes, which are precisely those that define an intrinsic Voronoi diagram on the features space, following the definitions given in the last section.

This geometric intuition matches the Fisher's discriminants theory, where for each pair of different distributions, the optimal decision threshold is precisely the place in the domain where the distributions get the same probability value. Following the reasoning, we can state now the relation among the intrinsic Bayes classifier and the Voronoi diagrams in the next theorem.

Theorem 3 (Bayes-Voronoi diagram) *Be M a differential manifold, $C\{c_k\}_{k\in K}$ a set of classes with a distribution defined on M, $P_C = \{\mu_k\}_{k\in K}$ a set of sites on M associated to each class, and $\mathcal{B}_g = \left\{\mathcal{B}_g^k(x)\right\}$ a family of Bayes distance functions associated to the classes. The, the tuple $(M, P_C, \mathcal{B}_g)$ defines a Voronoi diagram $\mathcal{V}_C$ (3.40), that we call Intrinsic Bayes-Voronoi diagram.*

$$\begin{aligned} \mathcal{V}_C &= (R_k)_{k\in K} \\ R_k &= \left\{x \in M : \mathcal{B}_g^k(x) \leq \mathcal{B}_g^j(x), \forall j \neq k\right\} \end{aligned} \tag{3.40}$$

The proof is trivial and it follows from the definition, because the functions $\mathcal{B}_g$ are local distances measured on the manifold M, and they are induced by the distributions for each class.

As future work, it would interesting to study the properties of the Voronoi diagrams to know if they can be helpful for the solution of our main problem.

3.4.5 Training of the model

The training of the model reduces to the estimation of the distribution$\mathcal{N}_g^k\left(\mu_k, \Sigma_g^k\right)$ for each class, problem that implies the estimation of the intrinsic covariance matrix Σ_g^k and its inverse. The covariance matrix is real-valued, symmetric and positive defined, and for the general case is not sparse. It implies that the computation of the inverse matrix has a $O\left(n^2\right)$ storing complexity and $O\left(n^3\right)$ computation complexity, therefore, the estimation is intractable for the common n values in the context of the document classification.

To overcome the described drawback, we propose as training method to find the distribution for each class on a specific subspace, what means to find the features that best describe the classes and project them on this subspace. Basically, we propose to use some features selection method. Using this approach we are approximating the covariance matrix Σ_g^k according to the number of selected features. The important issue is that we know where the solution "lives" and any better estimation of the distributions $\mathcal{N}_g^k$ move us toward it.

3.4.5.1 Model estimation

Starting from the training corpus D_T, we get the documents for a class c_k, denoted by $D_k = \{d \in D_T : \varphi(d) = c_k\}$. The training consists in the estimation of the priori probability $p(c_k)$ and the parameters of $\mathcal{N}_g^k\left(\mu_k, \Sigma_g^k\right)$ for each class c_k.

For this, we use the MLE method, which reduces to the expression (3.41) for the mean and (3.42) for the variances. As estimated $\widehat{p}(c_k)$ of the priori probability $p(c_k)$ we can use the distribution for the classes in the corpus.

Each document is represented by a vector $x_r \in S_+^n$.

$$\widehat{\mu}_k = \frac{1}{|D_k|}\sum_{r \in D_k} x_r \tag{3.41}$$

$$\begin{aligned}\widehat{\Sigma}_g^{k,ij} &= E\left[\left(g_\mu^T(x) \cdot \frac{\partial \mathcal{X}}{\partial u_i}\right)\left(g_\mu^T(x) \cdot \frac{\partial \mathcal{X}}{\partial u_j}\right)\right] \\ &= \frac{1}{|D_k|}\sum_{r \in D_k}\left(g_\mu^T(x_r) \cdot \frac{\partial \mathcal{X}}{\partial u_i}\right)\left(g_\mu^T(x_r) \cdot \frac{\partial \mathcal{X}}{\partial u_j}\right)\end{aligned} \tag{3.42}$$

For the parameters estimation, we can use all the vectors assigned to a same class, or we can use any other method to filter "outliers", such as the known random sampling (RANSAC).

3.4.5.2 Feature selection by class

As we mention, the computation of the inverse of the covariance matrix Σ_g^k is an intractable problem for the dimensions managed in the TC problem. In fact, the

estimation of the coefficients $\widehat{\Sigma}_g^{k,ij}$ is also hard, because it requires a storing in memory with quadratic complexity $O(n^2)$, preventing the possibility to represent the matrix in the main memory for any features space with large dimensions.

By other hand, we mentioned in the introduction about the VSM model that all the vectors use to be sparse. It is true in the Euclidean space, but due to the transformation of the problem to the tangent space, which depends on the position of the mean vector μ for each class, this property is not true for the general case.

A practical solution to estimate the model consists in the projection of the data on a lower dimensional space, using a combination of well established methods for features selection (χ^2) and dimensionality reduction (PCA [Zu et al., 2003]). If it would be desired, the first features selection method can be replaced by other alternative. However, for the second step, PCA is our favorite option because it is a closed and well founded algorithm, where the unique free parameter is the number of selected eigenvectors to define the final dimension of the transformed data.

If we select a set of representative coordinates (features) by class, we can define lower dimension unit hyperspheres $S_+^{n_k}$, with $n_k \ll n$. Using any features selection method, we can build the family of projections per class $\pi = \{\pi_k\}_{k \in K}$ in (3.43)

$$\pi_k : S_+^n \rightarrow S_+^{n_k} \tag{3.43}$$

As first step to classify any document, it is projected by the functions π_k on the subspaces $S_+^{n_k}$ associated to each class, then, the Bayes distance is computed on $S_+^{n_k}$. The optimal solution follows verifying the Bayes's criterion, but now, there are not a whole Voronoi diagram covering the whole features space, because the features spaces are individuals patches for each class.

The simplified optimal solution c^* is given by (3.44), where the distance functions $\widehat{\mathcal{B}}_g^k$ are computed using the normal distribution on the subspaces $S_+^{n_k}$.

$$c^* = \underset{c_k \in C}{argmin} \left\{ \widehat{\mathcal{B}}_g^k \circ \pi_k(x) \right\} \tag{3.44}$$

The definition of specific subspaces per class, denoted by $S_+^{n_k}$, allows the scalability of the method respect to the increasing of the classes, because each model for an individual class can be defined without impacts the rest of the classes. The subspaces for each class can be independently defined with its optimal number of features, without to need any trade-off about the total number of features of the VSM model, thing would happen if a global features space is defined, whose dimension needs to be increased every time a new class is added to the model.

The insertion of a novel class to the model only requires the selection of the set of descriptive features for the class and the estimation of its normal distribution. The only global parameter that is affected is the a priori probability for each class, which needs to be recomputed. This is a simple task if the system registers the number of documents classified for each class.

The proposed method to reduce the dimensionality per class is PCA, also known as Karhunen-Loeve transformation in signal processing, and Latent Semantic Index (LSI) in NLP.

Starting from the documents set for each class, denoted by D_k, we get the non-zero coordinates for all the vectors in the set, and we build a mapping function

for the coordinate indexes, to define the new vector $\widetilde{x}$. The number of non-null coordinates in D_k is denoted by n_k^r what defines the first projection (3.45) for the data of the class c_k. In this first step, we can also use any features selection method as the known chi-square χ^2.

$$\pi_k^r : S_+^n \to \mathbb{R}^{r_k} \tag{3.45}$$

Then, we define the covariance matrix $\Sigma_{k,r}$ (3.46) in the space $\mathbb{R}^{r_k}$, which has a dimension $r_k \times r_k$.

$$\Sigma_{k,r}^{ij} = E\left[\left(\widetilde{x}^i - \widetilde{\mu}_k^i\right)\left(\widetilde{x}^j - \widetilde{\mu}_k^j\right)\right],\ \widetilde{x}, \widetilde{\mu} \in \mathbb{R}^{r_k} \tag{3.46}$$

We get the SVD decomposition of the matrix $\Sigma_{k,r}$ (3.47) and we select the greater s singular values, getting an approximation of order s for the covariance matrix.

$$\Sigma_{k,r} = U \Lambda U^T \tag{3.47}$$

The covariance matrix is a *normal matrix* because it is square and verifies the condition $\Sigma_{k,r} \cdot (\Sigma_{k,r})^T = (\Sigma_{k,r})^T \cdot \Sigma_{k,r}$. Like consequence, the singular value decomposition (SVD) of the matrix matches the eigenvectors decomposition. The matrix Λ is diagonal and it contains the eigenvalues of $\Sigma_{k,r}$, while that the matrix U is orthogonal and it contains the eigenvectors of the covariance matrix. The matrices U and Λ has dimension $r \times r$.

Now, we can select the s_k greater eigenvalues and their associated eigenvectors, we get the sought transformation matrix π_s (3.48). This matrix defines the projection of a vector on the subspace spanned by the subset of eigenvectors of the covariance matrix, which are the directions in the Euclidean space that best represent the data for the approximation order s_k.

$$\begin{aligned} \pi_k^s &: \mathbb{R}^{r_k} \to S_+^{(s_k-1)} \\ \pi_k^s(x) &= \frac{1}{\left\|U_{s_k}^T \cdot x\right\|} U_{s_k}^T \cdot x \end{aligned} \tag{3.48}$$

Finally, we can define the whole projection associated to a class c_k by (3.49), which is composed by a features selection step π_k^r, followed by a dimensionality reduction transformation π_k^s based in PCA.

$$\begin{aligned} \pi_k &: S_+^n \to S_+^{n_k},\quad n_k = s_k - 1 \\ \pi_k &= \pi_k^s \circ \pi_k^r \end{aligned} \tag{3.49}$$

The training process requires the computation and storing of the projections π_k^s and π_k^r for each class, and the parameters for the intrinsic normal distribution $\mathcal{N}_g^k$ on the features subspace for each class. The training is summarized in the next steps:

1. Computation and storing of the terms and indexes for the r nonzero coordinates that define each projection π_k^r.

2. Computation and storing of the eigenvectors matrix $U_{s_k}^T$ what defines π_k^s.

3. Computation and storing of the parameters of $\mathcal{N}_g^k\left(\mu_k, \Sigma_g^k\right)$ on the features subspace $S_+^{n_k}$.

3.5 Experiments and discussion

To perform the validation of the proposed model, we repeated the experiments of text classification made by Lewis with the corpus RCV2 [Lewis et al., 2004]. We use the same weight vectors used in the benchmarks reported in this paper, with the aim to verify the model regardless of other factors, such as: the definition of the vocabulary, the feature selection method and the stemming method.

We use the weight vectors defined for the documents in the corpus, which had a size close to 50,000 and were organized into a hierarchy of 360 classes. We carried-out two different experiments: (1) n-single class binary classifiers (one per class $\rightarrow$ not global features space), and (2) a whole multiclass classifer on the same features space. Both experiments were implemented in C# .NET on a laptop with an Intel Pentium B950 @ 2.1 Ghz with 4 Gb RAM.

3.5.1 First experiment

Due to the large dimensionality of the problem, which impacts on the estimation of the covariance matrix, we implemented N independent binary classifiers, one per class, defined on the subspace of the features space containing all vectors of the same class. The classifiers worked in two stages: features selection (projection) + binary classification.

Simple features selection. To perform the classification of a vector, the feature vectors were projected onto the subspaces of each individual class. This projection process is equivalent to a features selection on the input vectors according to the particular features of each evaluated class, what basically produces a truncation of the components of input vectors that are not part of the subspace of each class, or in other words, the keywords that do not appear in the evaluated class are removed from the input vector.

During the implementation of the first experiments, we got the problems described below:

1. The number of training vectors per class was low and not evenly distributed, which led to singular covariance matrices and numerical difficulties for the model estimation. These problems prevented the estimation of the parameters of the model, such as it had been designed.

2. Most of similarity values (cosine function) between vectors was zero as a result of the huge spread of values and the sparse structure of the vectors, even within the same class, a condition that also greatly complicates the ability to adjust the parameters model.

3. The features selection per class, induced by the projection of the input vectors onto the subspace of each class, resulted in low rates of classification accuracy.

Conclusion. The above problems are summarized in that the structure of the feature vectors makes it impossible to adjust the proposed model, or what is the same, the proposed full model is not able to represent the observed data.

3.5.2 Second experiment

To solve the problems described above, we decided to make some changes in the model, trying to preserve the original spirit, in the hope to get better results. Following this idea, the following changes were made to the basic model:

1. **Independent components.** We assumed a model of independent components. This simplification of the model means that we only consider the variances of each individual feature, eliminating the cross covariance entries, what leading us therefore to a diagonal matrix with a trivial inverse. The resulting geometry for the distributions are quadrics aligned with the coordinate axes in the canonical Euclidean ambient space. We decided to assign a small random value (epsilon) to the off-diagonal values.

2. **Removal of the features selection.** To avoid the features selection induced by the projection on the subspace of each class, we decided to remove the first step of features selection, following the spirit of SVM, which is able to operate on vectors of large dimension without to use any dimensionality reduction technique. This led us to the following problems:

   (a) **High numerical and memory complexity.** Operate with vectors without reducing its dimension, leads to a huge increase in the computational complexity, as well as a runtime memory problem, as each vector is of order 5×10^4 components, what means around 4×10^5 bytes/vector. For a large training set, it is not possible to hold in the main memory the data in any standard computer.

   (b) **Impossibility to represent all tangent vectors in memory.** The computation of the geodesic vector involves the projection of each vector associated to a document ($\dim \sim 5 \times 10^4$) on the tangent space to the unit hypersphere, which involves the dot product of this vector with approximately 5×10^4 vectors de dimension $O(5 \times 10^4)$. The problem is that the 5×10^4 vectors defining the tangent space at the centroid of each class can not be represented in memory, because their storage would require $\sim 2 \times 10^{10}$ bytes.

3. **Solution with a high computational cost.** One possible solution to the problems (2.a) and (2.b), which was implemented at the expense of a high computational cost, is the computation at fly of the tangent vectors as they become necessary for their scalar product with the features vector to be classified. The sparse structure of the tangent vectors and the input vector allow the optimization of the computation of their scalar product, since it is only necessary to compute the product of the nonzero coordinates between the two vectors. Finally, this solution yielded poor precission results with ill-conditioned numerical stability.

Numerical ill-conditioning problem. Despite our effort to solve the practical problems of implementation of the proposed model in the second experiment, the last detected problem was the numerical instability of the parametrization of the unit hypersphere for dimensions as high as 5×10^4. The direct parameterization of the hypersphere requires the computation of a nested product of up to 5×10^4 values of the cosine function, while that the inverse function requires the computation of the square root for a sum of squares of the same order, which produced a huge numerical instability.

The model is not able to fit a typical keyword-based TC dataset, it is ill-conditioned numerically, and with a high computational cost if we discard the initial features selection (projections) step.

3.6 Conclusions

We have introduced a novel method for document classification based on the construction of a Bayes classifier on the feature space considered like a differential manifold. The classifier is optimal according to the geometry of the feature space and the normal distribution of the data.

We found a relationship between the optimal Bayes classifier and the geometric structure (decision boundaries) generated on the feature space, proving that said structure is an intrinsic Voronoi diagram which we call Bayes-Voronoi diagram.

We have also generalized some statistical objects on differentiable manifolds, such as the normal distribution and the Mahalanobis distance, and we have presented the Bayes distance. These objects are called respectively, geodesic Mahalanobis distance and Bayes distance.

Unfortunately, the proposed classifier model, although it was interesting from a theoretical point of view, the first experimental results indicate that it has an enormous difficulty of practical implementation. We attribute this problem to two possible causes: (1) the structure of the training data makes very difficult adjusting our generative model, and (2) the numerical complexity of the model itself, derived from the equations of the unit hypersphere, produces numerical instability for high-dimensional space features. Therefore, it is necessary to conduct a further review of the proposed model or leaving this research line.

Chapter 4

Final conclusions and future work

This thesis has introduced two novel semantic representation spaces for text documents and semantically annotated data, based in an intrinsic geometry approach, as well as other results, among which we have: a novel ontology-based semantic distance called *weighted Jiang-Conrath*, a generalized normal distribution on differential manifolds that we call *geodesic normal distribution*, which lead us to the definition of the *geodesic Mahalanobis distance.* By last, we prove that any Bayes classifier on a manifold defines a dual Voronoi diagram on its domain.

The ontology-based IR model looks promising, but it has not been evaluated experimentally yet, while that, the text classifier yielded some discouraging results.

Respect to the Intrinsic Bayes-Voronoi classifier, the model has yielded discouraging results as consequence of the difficulties for the training of the model, and the numerical instability derived from the numerical complexity of the hypersphere equations. By this reason, we have decided to suspend our research in this direction to focus in the more promising research trend about ontology-based IR models and semantic distances. Despite of these discouraging results, we could study the possibility to publish some partial results related with the Intrinsic Bayes-Voronoi classifier.

As next steps in our research, we can cite some different trends as follows. First, the development of some basic validation experiments for the *Intrinsic Ontological Spaces* model. Second, the extension of the experiments to large scale collections like the Web, and different domains. Third, the evaluation of the *Intrinsic Ontological Spaces* model in other NLP and IR tasks, such as the ontology-based clustering, word disambiguation, text summarization, automatic semantic annotation and Q&A among others. Fourth, the investigation of novel geometric search structures on semantic metric spaces based in the integration of the ontology in the search model.

The inquiry about ontology-based geometric search structures could open a new research trend that we would call *ontology-based space partitions* for semantically annotated data, which could play an important role for the semantic indexing of large data collections like the Web. This research trend would have as main objective the inquiry about the existence of ontology-based geometric search algorithms and structures, which could be used for very large scale semantic indexing and retrieval.

Like continuation of the work about manifold-based distributions, it would be interesting to browse of the literature about statistics differential geometry, field known as *directional statistics* [Mardia & Jupp, 2000], to study the definition of other dis-

tributions on manifolds, distinct from the geodesic normal distribution proposed here.

Bibliography

[Ahuja et al., 1990] Ahuja, R. K., Mehlhorn, K., Orlin, J., & Tarjan, R. E. (1990). Faster algorithms for the shortest path problem. *J. ACM*, 37(2), 213–223.

[Amari et al., 1987] Amari, S. I., Barndorff-Nielsen, O. E., Kass, R. E., Lauritzen, S. L., & Rao, C. R. (1987). Differential geometry in statistical inference. *Lect. Notes Monogr. Ser.*, 10, i–240.

[Apté et al., 1998] Apté, V, C., Damerau, F. J., & Weiss, S. M. (1998). Text mining with decision rules and decision trees. In *Proceedings of the Conference on Automated Learning and Discorery, Workshop 6: Learning from Text and the Web,*.

[Arandjelovic et al., 2005] Arandjelovic, O., Shakhnarovich, G., Fisher, J., Cipolla, R., & Darrell, T. (2005). Face recognition with image sets using manifold density divergence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1 (pp. 581–588).: IEEE.

[Arregui Fernández, 1988] Arregui Fernández, J. (1988). *Topología*. Madrid, España: Universidad Nacional de Educación a Distancia.

[Aurenhammer, 1991] Aurenhammer, F. (1991). Voronoi diagrams: a survey of a fundamental geometric data structure. *ACM Comput. Surv.*, 23, 345–405.

[Baker & McCallum, 1998] Baker, L. D. & McCallum, A. K. (1998). Distributional clustering of words for text classification. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 96–103). Melbourne, Australia: ACM.

[Basili & Pennacchiotti, 2010] Basili, R. & Pennacchiotti, M. (2010). Distributional lexical semantics: Toward uniform representation paradigms for advanced acquisition and processing tasks. *Natural Language Engineering*, 16, 347–358.

[Baumgart et al., 2006] Baumgart, M., Eckhardt, S., Griebsch, J., Kosub, S., & Nowak, J. (2006). *All-Pairs Common-Ancestor Problems in Weighted Dags*. Technical Report TUM-I0606, Institut f̀ur Informatik, Technische Universitat M'́unchen.

[Birkhoff, 1963] Birkhoff, G. (1963). *Lattice Theory*. Colloquium publications. American Mathematical Society.

[Birkhoff, 1967] Birkhoff, G. (1967). *Lattice Theory*, volume XXV of *Colloqium Publications*. American Mathematical Society, third edition.

[Boczko et al., 2009] Boczko, E. M., Minhui, X., Di, W., & Young, T. (2009). Comparison of binary classification based on signed distance functions with support vector machines. In *Bioinformatics, 2009. OCCBIO '09. Ohio Collaborative Conference on* (pp. 139–143).

[Boczko & Young, 2005] Boczko, E. M. & Young, T. R. (2005). The signed distance function: a new tool for binary classification. *Arxiv preprint cs/0511105*.

[Boissonnat & Karavelas, 2002] Boissonnat, J.-D. & Karavelas, M. I. (2002). *On the combinatorial complexity of Euclidean Voronoi cells and convex hulls of d-dimensional spheres*. Technical Report RR-4504, Sophia-Antipolis.

[Bratsas et al., 2007] Bratsas, C., Koutkias, V., Kaimakamis, E., Bamidis, P., & Maglaveras, N. (2007). Ontology-based vector space model and fuzzy query expansion to retrieve knowledge on medical computational problem solutions. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE* (pp. 3794–3797).: IEEE.

[Brin, 1995] Brin, S. (1995). Near neighbor search in large metric spaces. In *Proceedings of the 21st Conference on Very Large Databases (VLDB'95)* (pp. 574–584).: ilpubs.stanford.edu.

[Budanitsky & Hirst, 2001] Budanitsky, A. & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources*, volume 2.

[Budanitsky & Hirst, 2006] Budanitsky, A. & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist.*, 32, 13–47.

[Buhmann, 2003] Buhmann, M. (2003). *Radial basis functions: theory and implementations*, volume 12. Cambridge Univ Pr.

[Cai & He, 2012] Cai, D. & He, X. (2012). Manifold adaptive experimental design for text categorization. *IEEE Trans. Knowl. Data Eng.*

[Cao & Ngo, 2012] Cao, T. H. & Ngo, V. M. (2012). Semantic search by latent ontological features. *New Generation Computing*, 30, 53–71.

[Castells, 2008] Castells, P. (2008). Búsqueda semántica basada en el conocimiento del dominio. In F. Verdejo & A. García-Serrano (Eds.), *Acceso y visibilidad de la información en la red: el rol de la semántica* (pp. 111–138). España: Universidad Nacional de Educación a Distancia (UNED).

[Castells, 2013] Castells, P. (2013). Clarifications about the Castells-Fernández-Vallet ontology-based IR model. personal communication.

[Castells et al., 2007] Castells, P., Fernández, M., & Vallet, D. (2007). An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Trans. Knowl. Data Eng.*, 19(2), 261–272.

[Chai et al., 2002] Chai, K. M. A., Chieu, H. L., & Ng, H. T. (2002). Bayesian online classifiers for text classification and filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 97–104). Tampere, Finland: ACM.

[Chang & Lin, 2011] Chang, C. C. & Lin, C. J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2, 27.

[Chu-Carroll et al., 2012] Chu-Carroll, J., Fan, J., Boguraev, B., Carmel, D., Sheinwald, D., & Welty, C. (2012). Finding needles in the haystack: Search and candidate generation. *IBM J. Res. Dev.*, 56, 6: 1–6: 12.

[Chua & Kulathuramaiyer, 2004] Chua, S. & Kulathuramaiyer, N. (2004). Semantic feature selection using WordNet. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 166–172).: IEEE Computer Society.

[Clark, 2012] Clark, S. (2012). Vector space models of lexical meaning. In S. Lappin & C. Fox (Eds.), *Handbook of Contemporary Semantics*. Malden, MA: Blackwell, second edition.

[Clarke, 2007] Clarke, D. (2007). *Context-theoretic Semantics for Natural Language: an Algebraic Framework*. PhD thesis, University of Sussex.

[Clarke, 2009] Clarke, D. (2009). Context-theoretic semantics for natural language: an overview. In R. Basili & M. Pennacchiotti (Eds.), *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics* (pp. 112–119). Athens, Greece: Association for Computational Linguistics.

[Clarke, 2012] Clarke, D. (2012). A context-theoretic framework for compositionality in distributional semantics. *Comput. Linguist.*, 38, 41–71.

[Cohen, 1995] Cohen, W. W. (1995). Text categorization and relational learning. In *The Twelfth International Conference on Machine Learning (ICML'95).* (pp. 124–132).: Morgan Kaufmann Publishers, Inc.

[Cohen & Singer, 1999] Cohen, W. W. & Singer, Y. (1999). Context-sensitive learning methods for text categorization. *ACM Trans. Inf. Syst. Secur.*, 17, 141–173.

[Cortes & Vapnik, 1995] Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Mach. Learn.*, 20, 273–297.

[Cross et al., 2013] Cross, V., Yu, X., & Hu, X. (2013). Unifying ontological similarity measures: A theoretical and empirical investigation. *International Journal of Approximate Reasoning*, 54(7), 861–875.

[Dasgupta et al., 2007] Dasgupta, A., Drineas, P., Harb, B., Josifovski, V., & Mahoney, M. W. (2007). Feature selection methods for text classification. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 230–239). San Jose, California, USA: ACM.

[Deza & Deza, 2009] Deza, M. & Deza, E. (2009). *Encyclopedia of distances*. Springer.

[Dijkstra, 1959] Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numer. Math.*, 1(1), 269–271.

[Ding et al., 2007] Ding, L., Kolari, P., Ding, Z., & Avancha, S. (2007). Using ontologies in the semantic web: A survey. In R. Sharman, R. Kishore, & R. Ramesh (Eds.), *Ontologies*, volume 14 of *Integrated Series in Information Systems* (pp. 79–113).: Springer US.

[do Carmo, 1992] do Carmo, M. (1992). *Riemannian Geometry*. Boston: Birkhauser.

[Dragoni et al., 2010] Dragoni, M., Pereira, C. D. C., & Tettamanzi, A. G. (2010). An ontological representation of documents and queries for information retrieval systems. In *Trends in Applied Intelligent Systems* (pp. 555–564).: Springer.

[Dwyer, 1989] Dwyer, R. A. (1989). Higher-dimensional voronoi diagrams in linear expected time. In *Proceedings of the fifth annual symposium on Computational geometry* (pp. 326–333). Saarbruchen, West Germany: ACM.

[Egozi et al., 2011] Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)*, 29, 8.

[Eilenberg & MacLane, 1945] Eilenberg, S. & MacLane, S. (1945). General theory of natural equivalences. *Trans. Amer. Math. Soc.*, 58, 231–294.

[Erk, 2012] Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Lang. Linguist. Compass*, 6, 635–653.

[Erkan & Radev, 2004] Erkan, G. & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22, 457–479.

[Esuli et al., 2008] Esuli, A., Fagni, T., & Sebastiani, F. (2008). Boosting multi-label hierarchical text categorization. *Inf. Retr. Boston.*, 11(4), 287–313.

[Esuli & Sebastiani, 2009] Esuli, A. & Sebastiani, F. (2009). Active learning strategies for Multi-Label text classification. In *Advances in Information Retrieval*, Lecture Notes in Computer Science (pp. 102–113). Springer Berlin Heidelberg.

[Fang et al., 2005] Fang, W.-D., Zhang, L., Wang, Y.-X., & Dong, S.-B. (2005). Toward a semantic search engine based on ontologies. In *Proceedings of International Conference on Machine Learning and Cybernetics*, volume 3 (pp. 1913–1918).: IEEE.

[Feng et al., 2012] Feng, G., Guo, J., Jing, B.-Y., & Hao, L. (2012). A bayesian feature selection paradigm for text classification. *Inf. Process. Manag.*, 48(2), 283–302.

[Fernández et al., 2011] Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., & Motta, E. (2011). Semantically enhanced information retrieval: An ontology-based approach - JWS special issue on semantic search. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4), 434–452.

[Fernández et al., 2009] Fernández, M., López, V., Sabou, M., Uren, V., Vallet, D., Motta, E., & Castells, P. (2009). Using TREC for cross-comparison between classic IR and ontology-based search models at a web scale. In *Semantic Search 2009 Workshop at the 18th International World Wide Web Conference (WWW 2009),* Madrid, Spain.

[Fernández Laguna, 2003] Fernández Laguna, V. (2003). *Teoría básica de conjuntos.* Iniciación al método matemático. Madrid: Editorial Anaya.

[Fernández Sánchez, 2009] Fernández Sánchez, M. (2009). *Semantically enhanced Information Retrieval: an ontology-based approach.* PhD thesis, Universidad Autónoma de Madrid.

[Frakes & Baeza-Yates, 1992] Frakes, W. & Baeza-Yates, R. (1992). *Information Retrieval: Data Structures and Algorithms.* Prentice Hall PTR.

[Fresno, 2006] Fresno, V. (2006). *Representacion Autocontenida de Documentos HTML: una propuesta basada en Combinaciones Heurísicas de Criterios.* PhD thesis, Departamento de Ingeniería Telemática y Tecnologia Electrónica. Escuela Superior de Ciencias Experimentales y Tecnología. Universidad Rey Juan Carlos,.

[Fuhr et al., 1991] Fuhr, N., Hartmann, S., Knorz, G., Lustig, G., Schwantner, M., & Tzeras, K. (1991). AIR/X-a rule based multistage indexing system for large subject fields. In *Proceedings of RIAO'91* (pp. 606–623).

[Gabrilovich & Markovitch, 2006] Gabrilovich, E. & Markovitch, S. (2006). Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*, volume 6 (pp. 1301–1306). Boston, USA: AAAI Press.

[Gamboa & Ruiz, 2006] Gamboa, J. M. & Ruiz, J. M. (2006). *Introducción al estudio de las Variedades Diferenciales.* Editorial Sanz y Torres.

[Gan et al., 2013] Gan, M., Dou, X., & Jiang, R. (2013). From ontology to semantic similarity: calculation of ontology-based semantic similarity. *Scientific World Journal*, 2013, 11.

[García-Hernández & Ledeneva, 2009] García-Hernández, R. A. & Ledeneva, Y. (2009). Word sequence models for single text summarization. In *Second International Conference on Advances in Computer-Human Interactions (ACHI'09)* (pp. 44–48).: IEEE.

[Genkin et al., 2007] Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49, 291–304.

[Gold & Angel, 2006] Gold, C. & Angel, P. (2006). Voronoi hierarchies. In M. Raubal, H. Miller, A. Frank, & M. Goodchild (Eds.), *Geographic Information Science*, volume 4197 of *Lecture Notes in Computer Science* (pp. 99–111). Springer Berlin Heidelberg.

[Gupta & Gautam, 2014] Gupta, A. & Gautam, K. (2014). Semantic similarity measure using information content approach with depth for similarity calculation. *International Journal of Scientific & Technology Research*, 3(2), 165–169.

[Harispe et al., 2013] Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., & Montmain, J. (2013). A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. *J. Biomed. Inform.*

[Hatzivassiloglou et al., 2001] Hatzivassiloglou, V., Klavans, J. L., Holcombe, M. L., Barzilay, R., Kan, M.-Y., & McKeown, K. R. (2001). Simfinder: A flexible clustering tool for summarization. In *Proceedings of the NAACL workshop on automatic summarization* (pp. 41–49).

[He et al., 2008] He, C., Dong, Z., Li, R., & Zhong, Y. (2008). Dimensionality reduction for text using LLE. In *Natural Language Processing and Knowledge Engineering, 2008. NLP-KE '08. International Conference on* (pp. 1–7).

[Henrikson, 1999] Henrikson, J. (1999). Completeness and total boundedness of the hausdorff metric. *MIT Undergraduate Journal of Mathematics.*

[Hirst & St-Onge, 1998] Hirst, G. & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 305–332).: Massachusetts Institute of Technology.

[Hofmann et al., 2008] Hofmann, T., Sch\`olkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *Ann. Stat.*, (pp. 1171–1220).

[Hsieh et al., 2013] Hsieh, S.-L., Chang, W.-Y., Chen, C.-H., & Weng, Y.-C. (2013). Semantic similarity measures in the biomedical domain by leveraging a web search engine. *Biomedical and Health Informatics, IEEE Journal of*, 17(4), 853–861.

[Jiang & Conrath, 1997] Jiang, J. J. & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING X).*

[Joachims, 1998] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. N\'edellec & C. Rouveirol (Eds.), *Machine Learning: ECML-98*, volume 1398 of *Lecture Notes in Computer Science* (pp. 137–142).: Springer Berlin / Heidelberg.

[Kannan et al., 2012] Kannan, P., Bala, P. S., & Aghila, G. (2012). A comparative study of multimedia retrieval using ontology for semantic web. In *Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on* (pp. 400–405).

[Kass, 1989] Kass, R. E. (1989). The geometry of asymptotic inference. *Stat. Sci.*, (pp. 188–219).

[Khan et al., 2010] Khan, A., Baharudin, B., & Khan, K. (2010). Semantic based features selection and weighting method for text classification. In *International Symposium in Information Technology (ITSim'2010)*, volume 2 (pp. 850–855).

[KhounSiavash & Baraani-Dastjerdi, 2010] KhounSiavash, E. & Baraani-Dastjerdi, A. (2010). Using the whole structure of ontology for semantic relatedness measurement. *SEKE*.

[Klein, 1893] Klein, F. (1893). A comparative review of recent researches in geometry (translation the german paper published in erlangen, 1872). *Bull. Am. Math. Soc.*, (pp. 215–249).

[Koller & Sahami, 1997] Koller, D. & Sahami, M. (1997). Hierarchically classifying documents using very few words. In *The Fourteenth International Conference on Machine Learning (ICML'97)* (pp. 170–178).

[Kwok, 1998] Kwok, J. T.-Y. (1998). Automated text categorization using support vector machine. In *In Proceedings of the International Conference on Neural Information Processing (ICONIP*: Citeseer.

[Lam & Ho, 1998] Lam, W. & Ho, C. Y. (1998). Using a generalized instance set for automatic text categorization. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 81–89). Melbourne, Australia: ACM.

[Lastra Díaz & García Serrano, 2014] Lastra Díaz, J. J. & García Serrano, A. (2014). System and method for the indexing and retrieval of semantically annotated data using an ontology-based information retrieval model. *United States Patent and Trademark Office (USPTO) application*, US14/576,679.

[Leacock & Chodorow, 1998] Leacock, C. & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 265–283).: Massachusetts Institute of Technology.

[Lebanon, 2005] Lebanon, G. (2005). Information geometry, the embedding principle, and document classification. In *Proceedings of the 2nd International Symposium on Information Geometry and its Applications* (pp. 101–108).

[Lebanon, 2006a] Lebanon, G. (2006a). Metric learning for text documents. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28, 497–508.

[Lebanon, 2006b] Lebanon, G. (2006b). *Riemannian Geometry and Statistical Machine Learning*. PhD thesis, Carnegie Mellon University.

[Lebanon & Lafferty, 2004] Lebanon, G. & Lafferty, J. (2004). Hyperplane margin classifiers on the multinomial manifold. In *Proceedings of the twenty-first international conference on Machine learning* (pp.66̃). Banff, Alberta, Canada: ACM.

[Lee et al., 1993] Lee, J. H., Kim, M. H., & Lee, Y. J. (1993). Information retrieval based on conceptual distance in is-a hierarchies. *Journal of Documentation*, 49(2), 188–207.

[Lee et al., 2008] Lee, W.-N., Shah, N., Sundlass, K., & Musen, M. (2008). Comparison of ontology-based semantic-similarity measures. *AMIA Annu. Symp. Proc.*, (pp. 384–388).

[Lewis, 1998] Lewis, D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In *European Conference on Machine Learning: ECML-98* (pp. 4–15).

[Lewis & Ringuette, 1994] Lewis, D. D. & Ringuette, M. (1994). A comparison of two learning algorithms for text categorization. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94)*, volume 33 (pp. 81–93).

[Lewis et al., 2004] Lewis, D. D., Yang, Y., Rose, T. G., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5, 361–397.

[Lidl & Pilz, 1998] Lidl, R. & Pilz, G. (1998). *Applied Abstract Algebra*. New York: Springer-Verlag, second edition.

[Lin, 1998] Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning,*, volume 98 (pp. 296–304). Madison, WI.

[Lingling & Junzhong, 2012] Lingling, L. M. & Junzhong, J. G. (2012). A new model of information content based on concept's topology for measuring semantic similarity in WordNet1. *International Journal of Grid and Distributed Computing*, 5(3), 81–94.

[Lord et al., 2003] Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. A. (2003). Semantic similarity measures as tools for exploring the gene ontology. *Pac. Symp. Biocomput.*, (pp. 601–612).

[Machhour & Kassou, 2013] Machhour, H. & Kassou, I. (2013). Ontology integration approaches and its impact on text categorization. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 3, 31–42.

[Mardia & Jupp, 2000] Mardia, K. V. & Jupp, P. E. (2000). *Directional statistics*, volume 28. Wiley.

[Masand et al., 1992] Masand, B., Linoff, G., & Waltz, D. (1992). Classifying news stories using memory based reasoning. In *Proceedings of the 15th annual international ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 59–65). Copenhagen, Denmark: ACM.

[McCallum & Nigam, 1998] McCallum, A. & Nigam, K. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization,*, volume 752 (pp. 41–48).

[McKeown et al., 1999] McKeown, K. R., Klavans, J. L., Hatzivassiloglou, V., Barzilay, R., & Eskin, E. (1999). Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of the national conference on Artificial intelligence* (pp. 453–460).: John Wiley \& Sons Ltd.

[Meng et al., 2012] Meng, L., Gu, J., & Zhou, Z. (2012). A review of information content metric for semantic similarity. In *Advances on Digital Television and Wireless Multimedia Communications*, Communications in Computer and Information Science (pp. 299–306). Springer Berlin Heidelberg.

[Meng et al., 2005] Meng, W., Xiaorong, W., & Chao, X. (2005). An approach to concept-obtained text summarization. In *Communications and Information Technology, 2005. ISCIT 2005. IEEE International Symposium on*, volume 2 (pp. 1337–1340).

[Mihalcea & Tarau, 2004] Mihalcea, R. & Tarau, P. (2004). TextRank: Bringing order into texts. In *Proceedings of EMNLP*, volume 4: Barcelona, Spain.

[Miller, 1995] Miller, G. A. (1995). WordNet: A lexical database for english. *Commun. ACM*, 38, 39.

[Monjardet, 1981] Monjardet, B. (1981). Metrics on partially ordered sets: A survey. *Discrete Math.*, 35(1-3), 173–184.

[Moulinier, 1997] Moulinier, I. (1997). *Is learning bias an issue on the text categorization problem.* Technical report.

[Moulinier et al., 1996] Moulinier, I., Raskinis, G., & Ganascia, J. G. (1996). Text categorization: a symbolic approach. In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval* (pp. 87–99).

[Mouratis & Kotsiantis, 2009] Mouratis, T. & Kotsiantis, S. (2009). Increasing the accuracy of discriminative of multinomial bayesian classifier in text classification. In *Computer Sciences and Convergence Information Technology, 2009. ICCIT '09. Fourth International Conference on* (pp. 1246–1251).: ieeexplore.ieee.org.

[Munkres, 2000] Munkres, J. (2000). *Topology.* Prentice Hall, second edition.

[Mustafa et al., 2008] Mustafa, J., Khan, S., & Latif, K. (2008). Ontology based semantic information retrieval. In *Intelligent Systems, 2008. IS'08. 4th International IEEE Conference*, volume 3 (pp. 22–14–22–19). Varna: IEEE.

[Navigli, 2009] Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, 41, 10.

[Ng et al., 1997] Ng, H. T., Goh, W. B., & Low, K. L. (1997). Feature selection, perceptron learning, and a usability case study for text categorization. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 67–73). Philadelphia, Pennsylvania, United States: ACM.

[Orum & Joslyn, 2009] Orum, C. & Joslyn, C. A. (2009). Valuations and metrics on partially ordered sets.

[Pesquita et al., 2009] Pesquita, C., Faria, D., Falcao, A. O., Lord, P., & Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.*, 5, e1000443.

[Pierce, 1991] Pierce, B. C. (1991). *Basic Category Theory for Computer Science.* Cambridge, USA: Massachusetts Institute of Technology.

[Pirró & Euzenat, 2010] Pirró, G. & Euzenat, J. (2010). A feature and information theoretic framework for semantic similarity and relatedness. In *The Semantic Web Conference (ISWC' 2010)*, Lecture Notes in Computer Science (pp. 615–630). Springer Berlin Heidelberg.

[Pirró & Seco, 2008] Pirró, G. & Seco, N. (2008). Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content. In R. Meersman & Z. Tari (Eds.), *On the Move to Meaningful Internet Systems: OTM 2008*, volume 5332 of *Lecture Notes in Computer Science* (pp. 1271–1288).: Springer Berlin Heidelberg.

[Porter, 1980] Porter, M. (1980). An algorithm for suffix stripping. *Programmirovanie*, 14, 130–137.

[Rada et al., 1989] Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Trans. Syst. Man Cybern.*, 19(1), 17–30.

[Resnik, 1995] Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the IJCAI* (pp. 448–453).

[Rocchio, 1971] Rocchio, J. J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART Retrieval System: Experiments in Automatic Document Processing* (pp. 313–323.).: Prentice Hall.

[Roweis & Saul, 2000] Roweis, S. T. & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500), 2323–2326.

[Sahami, 1996] Sahami, M. (1996). Learning limited dependence bayesian classifiers. In *Proceedings of the Knowledge and Data Discovery Conference (KK'96)* (pp. 335–338).: AAAI.

[Salton et al., 1975] Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11), 613–620.

[Sánchez & Batet, 2012] Sánchez, D. & Batet, M. (2012). A new model to compute the information content of concepts from taxonomic knowledge. *Int. J. Semant. Web Inf. Syst.*

[Sánchez et al., 2011] Sánchez, D., Batet, M., & Isern, D. (2011). Ontology-based information content computation. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems*, 24(2), 297–303.

[Sánchez et al., 2012] Sánchez, D., Batet, M., Isern, D., & Valls, A. (2012). Ontology-based semantic similarity: A new feature-based approach. *Expert Syst. Appl.*, 39, 7718–7728.

[Sandler, 2005] Sandler, M. (2005). On the use of linear programming for unsupervised text classification. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 256–264). Chicago, Illinois, USA: ACM.

[Saruladha et al., 2011] Saruladha, K., Aghila, G., & others (2011). Information content based semantic similarity for cross ontological concepts. *International Journal of Engineering Science & Technology*, 3(6), 5132–5140.

[Saruladha et al., 2010] Saruladha, K., Aghila, G., & Raj, S. (2010). A survey of semantic similarity methods for ontology based information retrieval. In *Machine Learning and Computing (ICMLC), 2010 Second International Conference on* (pp. 297–301).: IEEE.

[Saruladha et al., 2012] Saruladha, K., Aghila, G., & Raj, S. (2012). Semantic similarity measures for information retrieval systems using ontology. In *Second International Conference on Machine Learning and Computing (ICMLC), 2010* (pp. 297–301).: IEEE.

[Schneider, 2005] Schneider, K. M. (2005). Techniques for improving the performance of naive bayes for text classification. In *Computational Linguistics and Intelligent Text Processing*, volume 3406 of *Lecture Notes in Computer Science* (pp. 682–693).: Springer.

[Sebastiani, 2002a] Sebastiani, F. (2002a). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34, 1–47.

[Sebastiani, 2002b] Sebastiani, F. (2002b). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34, 1–47.

[Sebastiani, 2005] Sebastiani, F. (2005). Text categorization. *Text mining and its applications to intelligence, CRM and knowledge management*, (pp. 109–129).

[Seco et al., 2004] Seco, N., Veale, T., & Hayes, J. (2004). An intrinsic information content metric for semantic similarity in WordNet. In R. López de M\ántaras & L. Saitta (Eds.), *Proceedings of the16th European Conference on Artificial Intelligence (ECAI)*, volume 16 (pp. 1089–1094). Valencia, Spain: IOS Press.

[Siddharthan et al., 2004] Siddharthan, A., Nenkova, A., & McKeown, K. (2004). Syntactic simplification for improving content selection in multi-document summarization. In *Proceedings of the 20th international conference on Computational Linguistics* (pp. 896).: Association for Computational Linguistics.

[Song & Tao, 2009] Song, D. & Tao, D. (2009). Discriminative geometry preserving projections. In *Proceedings of the 16th IEEE international conference on Image processing* (pp. 2429–2432). Cairo, Egypt: IEEE Press.

[Song & Tao, 2010] Song, D. & Tao, D. (2010). Biologically inspired feature manifold for scene classification. *IEEE Trans. Image Process.*, 19, 174–184.

[Stamatatos et al., 2000] Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Comput. Linguist.*, 26, 471–495.

[Su et al., 2008] Su, J., Zhang, H., Ling, C. X., & Matwin, S. (2008). Discriminative parameter learning for bayesian networks. In *Proceedings of the 25th international conference on Machine learning* (pp. 1016–1023). Helsinki, Finland: ACM.

[Swets & Weng, 1999] Swets, D. & Weng, J. (1999). Hierarchical discriminant analysis for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(5), 386–401.

[Taieb et al., 2012] Taieb, M. A. H., Ben Aouicha, M., Tmar, M., & Ben Hamadou, A. (2012). Wikipedia category graph and new intrinsic information content metric for word semantic relatedness measuring. In *Data and Knowledge Engineering*, Lecture Notes in Computer Science (pp. 128–140). Springer Berlin Heidelberg.

[Torkkola, 2004] Torkkola, K. (2004). Discriminative features for text document classification. *Pattern Anal. Appl.*, 6, 301–308.

[Turney & Pantel, 2010] Turney, P. D. & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res.*, 37, 141–188.

[Tzeras & Hartmann, 1993] Tzeras, K. & Hartmann, S. (1993). Automatic indexing based on bayesian inference networks. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 22–35). Pittsburgh, Pennsylvania, United States: ACM.

[Vallet et al., 2005] Vallet, D., Fernández, M., & Castells, P. (2005). An ontology-based information retrieval model. In *The Semantic Web: Research and Applications 2nd European Semantic Web Conference (ESWC 2005)* (pp. 455–470). Heraklion, Crete, Greece: Springer.

[Vanderwende et al., 2004] Vanderwende, L., Banko, M., & Menezes, A. (2004). Event-centric summary generation. *Working notes of DUC*.

[Wang & Chen, 2009] Wang, R. & Chen, X. (2009). Manifold discriminant analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009.* (pp. 429–436).

[Wang et al., 2012] Wang, Z., Sun, X., & Qian, X. (2012). Efficient kernel discriminative geometry preserving projection. *PRZEGLAD ELEKTROTECHNICZNY (Electrical Review)*, 5, 56–59.

[Wen et al., 2006] Wen, G., Chen, G., & Jiang, L. (2006). Performing text categorization on manifold. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics* (pp. 3872–3877).: IEEE.

[Widdows, 2004] Widdows, D. (2004). *Geometry and meaning*. CSLI publications Stanford.

[Wiener et al., 1995] Wiener, E., Pedersen, J. O., & Weigend, A. S. (1995). A neural network approach to topic spotting. In *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95)* (pp. 317–332).

[Wolf & Gibson, 2004] Wolf, F. & Gibson, E. (2004). Paragraph-, word-, and coherence-based approaches to sentence ranking: A comparison of algorithm and human performance. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (pp. 383).: Association for Computational Linguistics.

[Wu et al., 2011] Wu, J., Ilyas, I., & Weddell, G. (2011). *A study of ontology-based query expansion.* Technical Report CS-2011-04, University of Waterloo.

[Wu & Palmer, 1994] Wu, Z. & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94 (pp. 133–138). Stroudsburg, PA, USA: Association for Computational Linguistics.

[Xu & Shi, 2012] Xu, Q. & Shi, W. (2012). A comparison of semantic similarity models in evaluating concept similarity. In *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS'2012)*, volume XXXIX-B2, (pp. 173–178). Melbourne, Australia.

[Yan et al., 2007] Yan, S., Xu, D., Zhang, B., Zhang, H. J., Yang, Q., & Lin, S. (2007). Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29, 40–51.

[Yang, 1994] Yang, Y. (1994). Expert network: effective and efficient learning from human decisions in text categorization and retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 13–22). Dublin, Ireland: Springer-Verlag New York, Inc.

[Yang, 1999] Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Inf. Retr. Boston.*, 1, 69–90.

[Yang & Chute, 1994] Yang, Y. & Chute, C. G. (1994). An example-based mapping method for text categorization and retrieval. *ACM Transactions on Information Systems (TOIS)*, 12, 252–277.

[Yang & Liu, 1999] Yang, Y. & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the Twenty-Second Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 99)* (pp. 42–49).: ACM.

[Zhang et al., 2005] Zhang, D., Chen, X., & Lee, W. S. (2005). Text classification with kernels on the multinomial manifold. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 266–273). Salvador, Brazil: ACM.

[Zhou et al., 2008] Zhou, Z., Wang, Y., & Gu, J. (2008). A new model of information content for semantic similarity in WordNet. In *Future Generation Communication and Networking Symposia, 2008. FGCNS'08. Second International Conference on*, volume 3 (pp. 85–89).: IEEE.

[Zu et al., 2003] Zu, G., Ohyama, W., Wakabayashi, T., & Kimura, F. (2003). Accuracy improvement of automatic text classification based on feature transformation. In *Proceedings of the 2003 ACM symposium on Document engineering* (pp. 118–120). Grenoble, France: ACM.