DOKTORI (Ph.D.) ÉRTEKEZÉS

EFFICIENT METHODS IN THE PRACTICE OF INFORMATION RETRIEVAL

HATÉKONY MÓDSZEREK AZ INFORMÁCIÓ-VISSZAKERESÉS GYAKORLATÁBAN

Góth Júlia

Témavezető: Dr. Dominich Sándor

Veszprémi Egyetem
Műszaki Informatikai Kar
Informatikai Tudományok Doktori Iskola

2005

TARTALMI KIVONAT

Az információ-visszakeresésben a dokumentumokat indexkifejezésekkel reprezentálják, amelyeket vagy automatikusan a dokumentumból nyernek, vagy szakértők határoznak meg manuálisan. Attól függően, hogy melyik módszer használatos indexkifejezések megállapítására, más és más hatékonyságú lesz az információ-visszakereső rendszer. Szerző bevezet egy olyan eljárást indexkifejezések meghatározására és súlyszámításra, amely szélesebb körben használható és hatékonyabb visszakereső rendszert eredményez, mint az addigi hagyományos eljárás.

Az információ-visszakereső rendszer által visszaadott dokumentumok relevanciájukban eltérnek egymástól. Ez a felhasználó számára fontos információ, ugyanis befolyásolja abban, hogy mely válaszokat és azokat milyen sorrendben nézze meg. Egy információ-visszakereső rendszer kategoricitási tulajdonsága azt mutatja, hogy a rendszer a válaszaiban mennyire kategorikus, azaz a visszaadott dokumentumok relevanciájukban egymástól mennyire különböznek. A kategoricitást lehet változtatni, azonban ez "költséges" (nagy számításigényű) eljárás. Mivel a kategoricitás — felhasználói szempontból — a rendszer fontos tulajdonsága, célszerű ennek változtatását minél alacsonyabb költségű módszerrel megvalósítani. Szerző megad egy új, kisebb számítási bonyolultágú eljárást kategoricitás változtatására, amelyhez egy új — hiperbolikus geometrián alapuló — információ-visszakereső modellt vezet be.

ABSTRACT

In Information Retrieval (IR), the documents are represented by index terms created manually or automatically. The effectiveness of information retrieval system depends greatly on how the index terms are created. In the dissertation, a new method is proposed for the computation of term discrimination values, which presents advantages over the traditional vector-based calculation: it is faster and its application is not restricted to the Vector Space Model (VSM).

The key goal of an IR system is to retrieve information for a given query, which might be useful or relevant to the user. The returned answers differ from one another in their relevance values. A new concept called retrieval "categoricity"is introduced in the dissertation, which means the spreading of the answers' relevance values. Categoricity can be varied in the traditional VSM model, but it is a costly process. In the dissertation, a new and efficient way is proposed to vary retrieval categoricity using a new information retrieval technique based on hyperbolic geometry.

ABSTRAKT

In der Informationswiedergewinnung *(information retrieval)* werden Dokumente durch manuell oder automatisch erstellte Schlüsselwörter (*index terms*) repräsentiert. Die Methode der Erstellung der Indexterme hat einen wesentlichen Einfluss auf die Effektivität des Informationswiedergewinnungssystems. In dieser Arbeit wird ein neuer Ansatz für die Bestimmung der Schlüsselwörter und die Berechnung derer Gewichte vorgeschlagen, welcher effizienter als die traditionelle, auf Vektoren basierende Berechnung ist und dessen Anwendung sich nicht nur auf das Vektorraummodell beschränkt.

Für den Benutzer eines Informationswiedergewinnungsystems haben die als Ergebnis einer Abfrage zurückgelieferten Dokumente unterschiedliche Relevanzwerte. Die Kategorizität, die in der Dissertation eingeführte Eigenschaft des Informationswiedergewinnungsystems, gibt an, in welchem Ausmaß sich die Relevanzwerte der Abfrageergebnisse voneinander unterscheiden. Man kann die Kategorizität eines Informationswiedergewinnungssystems zwar anpassen, in dem traditionellen Vektorraummodell ist das aber ein rechenaufwändiges Vorgehen. Der in dieser Arbeit vorgestellte, neue, auf hyperbolischer Geometrie basierende Informationswiedergewinnungsansatz ermöglicht eine effizientere Anpassung der Systemkategorizität.

CONTENTS

ACKNOWLEDGEMENT

This work would never have been written without the help and continuous support of several people. First of all, I want to express my sincere gratitude to my supervisor, Sándor Dominich for his excellent and patient guidance throughout my work.

Very special thanks must go to Professor Ferenc Friedler for the highly supportive environment provided at the Department of Computer Science, University of Veszprém.

Many thanks go to my colleagues Tamás Kiezer, Adrienn Skrop, Zoltán Szlávik for the joint work, and for their helpful comments and suggestions in the period of writing this thesis. Furthermore thanks are due all the people in the Department of Computer Science.

I am very grateful to my husband, my parents and all my family for supporting my work in many ways for such a long time.

The support of the following project is kindly acknowledged: National Research and Development Fund of the Széchenyi Plan, Hungary, Consortium NKFP OM #2/052/2001, grants AKP #2001-140, and OTKA #T030194.

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Motivations

Recently, Information Retrieval (IR) has become one of the most important theoretical and practical research topics in information and computer science. Information retrieval deals with the representation, storage, organization of, and access to information items (Baeza-Yates and Ribeiro-Neto, 1999). The representation and organization of information should provide the user with easy access to the information in which he/she is interested. Essentially, IR means that there is a set of documents (or objects), and a person (user) asks a question (query) to which the answer is a set of relevant documents satisfying the information need expressed by his/her question. In IR several different models have been elaborated differing from each other in the way objects (documents and queries) are represented and in which retrieval itself is modelled. One of the most important, well-understood and extensively researched classical models is the Vector Space Model (VSM). (Baeza-Yates and Ribeiro-Neto, 1999; Dominich, 2001; Salton, 1966; Van Rijsbergen, 1979, 1987).

Note. Looking for information in natural language texts based on a query is obviously also a linguistic problem to a large extent. Such a problem raises many questions (e.g., reprensentation of the language space, what is information?, etc.) that go beyond the area of computer science per se. However, solution have been found that, of course, are not complete but make a computation-like treatment possible (for example, the Vector Space Model).

In information retrieval, the documents are represented by index terms created manually or automatically. The effectiveness of the information retrieval system depends, to a great extent, on how the index terms are created. The Term Discrimination Model (TDM) was introduced in (Salton, Yang, and Yu, 1974; Salton, Yang, and Yu, 1975) as a contribution to the automatic indexing theory in the Vector Space Model of information retrieval. I propose a new method for the computation of term discrimination values, which presents advantages over the traditional vector-based calculation. It is faster and its application is not restricted to the Vector Space Model.

The key goal of an IR system is to retrieve information for a given query, which might be useful or relevant to the user. The returned answers also (called hits) differ from each other in their relevance values for a given query. A new concept called retrieval "categoricity" (i.e., how categorical the hits are) was introduced in my dissertation, which means the spreading of the answers' relevance values. Categoricity can be varied in the traditional VSM model, but it is a costly process. I introduce a new and efficient way to vary retrieval categoricity using a new information retrieval technique based on hyperbolic geometry.

1.2 Contributions

The dissertation consists of 6 chapters. C*hapter 1* is an introduction containing the motivation and the contribution of the dissertation, and it gives a brief literature overview of the most important books in information retrieval. *Chapter 2* shows the methods applied and test collections used in my dissertation. Two standard test collections — named ADI and Reuters — further a Belief database including Hungarian belief texts as well as a Medical database developed by our research group (CIR = Center for Information Retrieval), are used in my experiments.

The new results obtained in my research are presented in the forthcoming three chapters. Each chapter begins with the description of the motivation behind that research. This is followed by a presentation of the results needed for understanding the forthcoming sections. The layout of the dissertation and the main scientific contributions are described below:

Chapter 3 investigates the possibility of defining a VSM in a hyperbolic space. In general, Euclidean geometry is the only type of space used in the VSM. In information retrieval, non-Euclidean spaces are used for information visualisation (Phillips and Gunn, 1992; Phillips, Levy, and Munzner, 1993). In section 3.4, I introduce the HIR (Hyperbolic Information Retrieval) Model; the similarity measure is derived from the Cayley-Klein hyperbolic distance. In section 3.5, it is shown formally as well as experimentally — using Medical Database — that the HIR model is equivalent to the Cosine-based VSM using normalised weighting scheme. The application called NeuRadIR is also presented, which is the first application using the HIR model.

In *Chapter 4*, I investigate the retrieval categoricity of the VSM and the HIR model. Then, I introduce a new efficient way to vary retrieval categoricity. The concept of entropy is used to define an amount of uncertainty *U* associated with answers in the Vector Space Model of information retrieval, and to define the connected concept categoricity. In section 4.4, it is shown that any retrieval model or system based on positive RSV (Retrieval Status Value) may be conceived as a probability space that decreases the amount of the associated Shannon information**.** In section 4.6, I investigate the retrieval categoricity of the VSM — using different similarity measures and weighting schemes — and it is shown experimentally that in the VSM the only way to modify the retrieval categoricity is to take a different weighting scheme and/or similarity measure. Thus, the Cosine measure with a *tfn* weighting scheme — one of the most commonly used methods — is the least categorical in its answers. Therefore, it is not enough to change only the weighting scheme or the similarity measure but both of them need to vary to obtain a more categorical sytem. This in turn yields costly re-computation of both weights and similarity measure values; while the same set of answers containing the same document with the same order cannot be guaranteed. In section 4.7, it is shown experimentally that in HIR retrieval categoricity depends only on the radius of the space. Thus, increasing the radius of the hyperbolic space yields a less categorical retrieval system and conversely: decreasing the radius leads to more categorical answers. It is shown in section 4.8 that in HIR a modifiable categoricity can be obtained at much lower re-computation costs: only the similarity values need be re-computed but not the weights, while rank order is preserved.

In *Chapter 5*, the concept of UDO (Uncertainty Decreasing Operation) — defined in section 4.3 — is proposed as a theoretical background for term discrimination power and it is applied to the computation of term discrimination values. Experimental evidence is given as regards such computation; the results obtained compare well to those obtained using vector-based calculation of term discrimination values. It is shown, that the UDO-based computation, however, presents advantages over the vector-based calculation: it is faster (section 5.4), easier to assess and handle in practice and its application is not restricted to the Vector Space Model, but it can be used in any positive RSV-based information retrieval system (section 5.3).

Chapter 6 gives a summary of the results obtained.

1.3 Literature Overview

Since the 1940s the problem of information storage and retrieval has become increasingly important: there are huge amounts of information to which accurate and speedy access is becoming more difficult (Van Rijsbergen,1979). The key goal of an information retrieval system is to retrieve information for a given query, which might be useful or relevant to the user.

In the field of information retrieval a huge number of articles have been published in specialized journals (e.g. Information Retrieval, Information Processing and Management, Journal of the American Society for Information Science) and at conferences (ACM SIGIR, ECIR) dealing with many different aspects of IR. A number of books have also been written about IR with a broad (and extensive) coverage of the various topics in the field, for example: Van Rijsbergen, 1975; Salton and MacGill, 1983; Kowalski, 1997; Baeza-Yates and Ribeiro-Neto, 1999; Dominich, 2001:

"Information Retrieval" (C. J. Van Rijsbergen, 1979) can be considered as a reference book of this field. There is a Hungarian translation dated 1987. The aim of this book was to give a complete coverage of the most important ideas in various special areas of information retrieval. The major change in the second edition of this book is the addition of a new chapter on probabilistic retrieval.

"Information Retrieval Systems: Theory and Implementation" (G. Kowalski, 1997) provides a theoretical and practical explanation of the latest advancements in information retrieval and their application to existing systems. It takes a systems approach, and presents all aspects of an information retrieval system. The importance of the Internet and its associated hypertext-linked structure, the human interface, and the importance of information visualization for identification of relevant information are also discussed.

"Modern Information Retrieval" (R. Baeza-Yates, and B. Ribeiro-Neto, 1999) presents an overall view of research in IR from a computer scientist's perspective. This means that the focus of the book is on computer algorithms and techniques used in information retrieval systems, and on trying to understand how people interpret and use information as opposed to how to structure, store, and retrieve information automatically. Most of this book is dedicated to the computer scientist's viewpoint of the IR problem; the human-centred viewpoint is discussed to some extent in the last two chapters. Additionally, this book puts a great emphasis on the integration of the different areas, which are closely related to the information retrieval problem, and thus should be treated together. For that reason this book also discusses visualization, multimedia retrieval and digital libraries.

"Mathematical Foundations of Information Retrieval" (S. Dominich, 2001) gives formal mathematical descriptions of the retrievals in the basic IR models in a unified mathematical style, format, language, and it creates an axiomatic, consistent mathematical framework.

"The Geometry of Information Retrieval" (C.J. van Rijsbergen, 2004) is an attempt to create a general formal "language" for basic IR models (coordination level matching, vector space, probabilistic, ostensive) using the concept of the Hilbert space.

The most important results — published in journal articles or conference proceedings — connected to the results of the dissertation are presented in the corresponding sections.

CHAPTER 2

METHODS APPLIED AND TEST COLLECTIONS USED IN EXPERIMENTS OF MY DISSERTATION

2.1 Retrieval Evaluation

Beyond so much success, the Web has introduced new problems of its own. Finding useful information on the Web is frequently a tiring and difficult task. For instance, to satisfy an information need, the user might navigate the space of Web links searching for information of interest. However, since the Web is huge and almost unknown, such a navigation task is usually inefficient. For naive users, the problem becomes harder, which might entirely frustrate all their efforts. The main difficulty is the semi-structured data model for the Web, which implies that information definition and structure are frequently of low quality. These difficulties have increased interest in IR and in its techniques as promising solutions. As a result, IR has gained a place with other technologies at the centre of the stage (Baeza-Yates and Ribeiro-Neto, 1999).

Relevance is the concept on which the whole theory and practice of IR is based. Relevance is a complex and widely studied idea in several fields from philosophy to library science; so far it plays an important role not only in information retrieval, but e.g. in information science, too.

Information retrieval systems require the evaluation of how precise the set of answers is. The evaluation measure for a given retrieval strategy quantifies the similarity between the set of documents retrieved and the set of relevant documents provided by the specialist. Simply, it points the goodness of the retrieval strategy. The following traditional measures are used to express how well (or badly) an IR system performs (*Ret* denotes the set of retrieved documents, and *Rel* denotes the set of relevant documents, and |.| denotes cardinality):

- *Precision:* is defined as the ratio of the number of relevant and retrieved documents to the total number of documents retrieved (it shows the fraction of the retrieved documents which is relevant):

$$Precision = \frac{|Ret \cap Rel|}{|Ret|}$$

- *Recall:* is defined as the ratio of the number of relevant and retrieved documents to the total number of relevant documents (it shows fraction of the relevant documents which has been retrieved):

$$Recall = \frac{|Ret \cap Rel|}{|Rel|}$$

- *Fallout*: is defined as the proportion of relevant and non-retrieved documents to the number of non-retrieved documents (it shows which fraction of the non-retrieved documents are relevant):

$$Fallout = \frac{|\overline{Ret} \cap Rel|}{|\overline{Ret}|}$$

There is a relationship between all three measures via a parameter called *generality* (*G*), which is a measure of the density of relevant documents in the collection. The formula is:

$$Precision = \frac{Recall \cdot G}{Recall \cdot G + Fallout \cdot (1 - G)},$$

G = $\frac{|Rel|}{N}$, where *N* denotes the number of documents

There are many studies and book chapters (e.g. in C. J. Van Rijsbergen, 1979; Salton and MacGill, 1983; Kowalski, 1997; Baeza-Yates; and Ribeiro-Neto, 1999; Belew, 2000; Dominich, 2001.) on evaluating the effectiveness of a given retrieval strategy.

In the evaluation of Web search engines recall is impossible to be evaluated. (Oppenheim, Morris and McKnight, 2000). However methods have been suggested to evaluate recall (Gordon and Pathak, 1999; Chu and Rosenthal,1996). (Leighton and Srivastava, 1999) elaborated and applied a method to evaluate and compare five Web search engines (Alta Vista, Excite, HotBot, Infoseek, Lycos) for precision on the first twenty results returned for fifteen queries. This method is described in details in section 3.6.2.

2.2 Text Collections Used

The evaluation in information retrieval usually based on a reference test collection and on an evaluation measure is called retrieval performance evaluation. Accordingly, in my dissertation test collections are used for experiments and demonstrations. A test collection consists of a collection of documents, a set of information requests, and a set of relevant documents for each information request. Many test collections have been created (Spark Jones, and van Rijsbergen, 1976).

A few of these collections are freely available on the web (http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/ or http://www.cs.utk.edu/~lsi/corpa.html) and are used by many researchers in information retrieval. The most popular standard test collections are: TREC, ADI, MED, CACM, CISI, TIME, REUTERS. These collections vary in size, topic and in the number of queries. Two popular and well-studied standard test collections (ADI and REUTERS) are used in my dissertation for my experiments.

2.2.1 ADI Test Collection

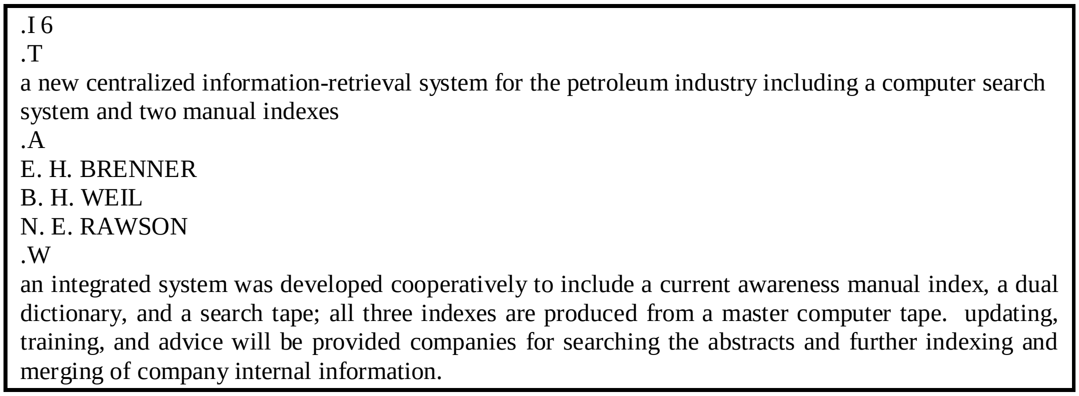
The ADI collection (http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/) is the smallest; it contains 82 homogeneous English articles from computing journals with 2086 index terms and 35 queries. The test collection contains the following files:

- adi.all : documents,
- adi.que : queries,
- adi.rel : relevance assessments,
- adi.bln : list of Boolean queries.

Documents (adi.all) contain text articles and additionally some structured fields. Figure 2.1 shows the 6th document of the collection for illustration. The documents generally include the following fields:

- .I : serial number of the document
- .T: title of the document
- .A: author/authors of the document
- .W: text of the document.

When the author/authors or other field of the articles is unknown, the corresponding field is missing from the document.



```
.I 6
.T
a new centralized information-retrieval system for the petroleum industry including a computer search
system and two manual indexes
.A
E. H. BRENNER
B. H. WEIL
N. E. RAWSON
.W
an integrated system was developed cooperatively to include a current awareness manual index, a dual
dictionary, and a search tape; all three indexes are produced from a master computer tape.  updating,
training, and advice will be provided companies for searching the abstracts and further indexing and
merging of company internal information.
```

Figure 2.1. Structure of the .I6 document

The ADI test collection contains 35 queries (adi.que). The structure of the query is similar to the documents (figure 2.2); implicitly it contains only the "serial number" (I), and the "text" (W) fields. Figure 2.2 illustrates the types of queries; .I3 is a mixed query including question and declarative sentence too, .I4 is a simple query including a declarative sentence and .I14 is a simple query including only a question.

```
.I 3
.W
What is information science?  Give definitions where possible.

.I 4
.W
Image recognition and any other methods of automatically transforming printed text into computer-
ready form.

.I 14
.W
What future is there for automatic medical diagnosis?
```

Figure 2.2. Structures of the .I3, .I4, .I14 queries

The information retrieval systems cannot process natural language queries; therefore it is necessary to process the queries to the appropriate form (e.g. to Boolean form in case of Boolean retrieval system). Figure 2.3 shows the Boolean forms of the queries in Figure 2.2.

```
#q3= #and ('information',#or ('science', 'definition'));

#q4= #or (#and ('image', 'recognition'), #and ( #or ('printed', 'text'),  #or ('methods', 'automatically',
        'transforming', 'computer-ready') ) );

#q14= #and ('medical', #or ('future', 'automatic')) ;
```

Figure 2.3. Boolean form of the I.3, I.4, I.14 queries

The ADI test collection includes the relevance assessments (adi.rel). Figure 2.4 shows a fraction of this file illustrating which answers are relevant to which queries (documents .I3, .I43, .I45, .I60 are relevant to the query .I3).

3	3
3	43
3	45
3	60
4	29
4	63
14	20
14	33

Figure 2.4. Relevance assessments to the query I.3, I.4, I.14

In my experiments, terms were selected from the ADI documents (number of term was 915). These documents were TIME stoplisted (http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/time/) and — to reduce a word to its *stem* or root form, thus, to represent the key terms of a query or document by stems instead of the original words — Porter stemmed (http://www.tartarus.org/~martin/PorterStemmer/).

2.2.2 Reuters Database

Reuters-21578 text categorisation test collection (http://www.research.att.com/~lewis) is another standard test collection; a part of it was used in my dissertation. It is a resource for information retrieval, machine learning and other corpus-based research. It is a public collection.

The documents in the Reuters-21578 collection appeared on the Reuters newswire in 1987. The documents were assembled and indexed with categories by personnel from Reuters Ltd. (Sam Dobbins, Mike Topliss, Steve Weinstein) and Carnegie Group, Inc. (Peggy Andersen, Monica Cellio, Phil Hayes, Laura Knecht, Irene Nirenburg) in 1987.

In 1990, Reuters and CGI made the documents available for research purposes to the Information retrieval Laboratory (W. Bruce Croft, Director) of the Computer and Information Science Department at the University of Massachusetts at Amherst. David D did formatting of the documents and production of associated data files in 1990. Lewis and Stephen Harding at the Information Retrieval Laboratory. David D. Lewis and Peter Shoemaker at the Center for Information and Language Studies, University of Chicago did further formatting and data file production in 1991 and 1992. This version of the data was made available for anonymous FTP as "Reuters-22173, Distribution 1.0" in January 1993. From 1993 through 1996, Distribution 1.0 was hosted at a succession of FTP sites maintained by the Center for Intelligent Information retrieval (W. Bruce Croft, Director) of the Computer Science Department at the University of Massachusetts at Amherst. At the ACM SIGIR '96 conference in August 1996 a group of text categorisation researchers discussed how published results on Reuters-22173 could be made more comparable across studies. They decided on producing a new version of collection with less ambiguous formatting, and including documentation carefully spelling out standard methods of using the collection. The opportunity would also be used to correct a variety of typographical and other errors in the categorization and formatting of the collection.

Steve Finch and David D. Lewis did this cleanup of the collection September through November of 1996, relying heavily on Finch's SGML-tagged version of the collection from an earlier study. One result of the re-examination of the collection was the removal of 595 documents, which were exact duplicates (based on identity of timestamps down to the second) of other documents in the collection. The new collection therefore has only 21,578 documents, thus called the Reuters-21578 collection.

The Reuters-21578 collection is distributed in 22 files. Each of the first 21 files (reut2-000.sgm through reut2-020.sgm) contains 1000 documents, while the last (reut2-021.sgm) contains 578 documents. The files are in SGML format.

In my experiments, only a part of the Reuters Database was used with 7000 documents and 32589 index terms. The text pre-processing was carried out automatically. After the automatic removal of the stop words using the TIME stop list (http://www.dcs.gla.ac.uk/idom/ir_resources/test_collections/time/), terms were generated automatically, without stemming. Stemming was not necessary for these experiments, because the database was not being used for retrieval. The average number of terms per document was 73.

2.2.3 Belief Database

Currently, there is no standard test collection containing Hungarian texts, but several scientific fields created their own text collections.

One example is the Belief Database including 2704 Hungarian belief texts, (http://www.oszk.hu/hun/publ/konferencia/konf2001/finn/finn_daranyi_hu.htm). This collection was also used in my experiments.

This collection has several specialities:

- it is not a standard one,
- it was special in using different spellings: contemporary Hungarian; for instance, "*Ha kis gyermeknek komoly baja van, akkor szenes vizzel mossák meg. A meleg vizbe 9 drb. szenet tesznek, megkenik a vizzel a gyermek homlokát és ezt mondják: Ha férfi, kalap alá; ha leány párta alá; ha asszony fejkötô alá, az atya, fiú, szentlélek nevében. Amen.*"
- a mixture of older Hungarian spelling and dialect was used; for example, "*Ha a tehenet merrontya a boszorkány, vësznek egy új fëlliteres cserepbëgrét; abba belëtësznek ecs csomaócskát a tehen gannajjábó. Azután szöget vernek a kény belsejébe s erre felakasztyák a bëgrét. Etteô aszt meggyön a tehen haszna.*"
- many different word forms were used.

Due to these characteristics the text pre-processing operations were carried out manually. A number of 1,551 stop words (e.g., pronouns, adverbs, articles, attributes, verbs, present participles, rarely used chemical words, as well as conjugated/declined forms) were identified manually as baring no or very little significance for beliefs, and gathered in a list. For example, the personal pronoun "*aki*", meaning "who", has many different declined forms such as: "*aki, akié, akiébe, akiért, akihez, akijé, akik, akiknek, akin, akinek, akinél, akire, akiről, akit, akitől, akivel, akki, akkinek, akkire*". After the automatic removal of the stop words there remained 14,286 word forms. The word forms were then stemmed manually. For example, the following declined word forms: "*csont, csontjával, csontja, csontig, csontok, csontjait, csontnak, csontokat, csontra, csontjai, csontom, csont, csonton, csontját, csontokbúl, csontot, csonttal*" were all stemmed to "*csont*" meaning "bone". A further difficulty stemmed from the very many composed words, which are typical of the Hungarian language (just like in German or Finnish, for instance). A further and very special difficulty was posed by old synonym words, which are not being used anymore in contemporary Hungarian; for example, the words "*betyöleges, bíszbányosok*" were replaced by the word "*varázs*" meaning "magic".

The result was a number of 2,602 terms in a correct contemporary Hungarian spelling, which were used as index terms for the belief texts. The average number of terms per text was 15.

2.2.4 Medical Database

A medical database[1], — which was developed in the Department of Computer Science within the Cost Effective Health Preservation Consortium Project —, was used for some experiments in *Chapter 3*, and *4*. This medical database is a part of the system called NeuRadIR. NeuRadIR is a **Neu**ro**Rad**iological **I**nformation **R**etrieval system to brain CT image and report retrieval. It was developed by our center, the CIR (Center for Information Retrieval). The implemented system enables physicians (both radiologist and general practitioners) to use medical text and image database over the Web in order to facilitate health preservation but also to assist diagnosis and patient care. The details of the system can be found in section 3.5.1.

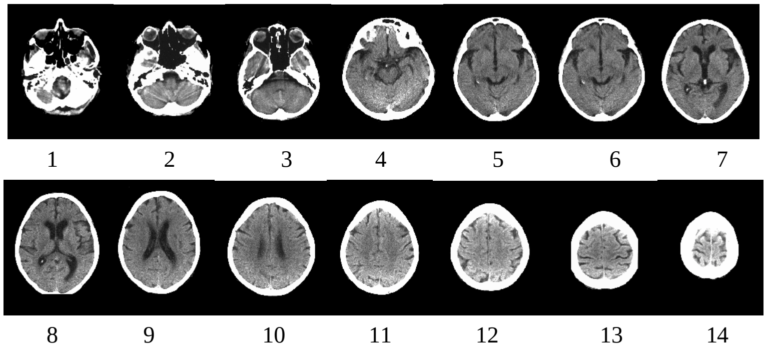
The medical database contains 40 medical cases denoted by a number (1-20) and a letter (a, or b). Every number has "a" and "b" version meaning that an early (denotes by "a"), and a late (denotes by "b") examination belong to every case. The database was created in English using ACTILYSE program.

Each case includes two parts (figure 2.5 shows an example):

(i) Computer Tomography (CT) images of the human patients' brains (each case contains from 10 to 14 image slices),

(ii) Textual information (scanning time, patient age, patient gender, patient notes, paresis information) and case report (the demographic data in this paper is not real to ensure anonymity).

Slices (1 through 14: cross sections from the bottom to the top of the head)



Patient: 70-year old woman

Patient notes: She was not aphasic and was fully conscious. The patient had a severe left-sided hemiparesis.

Case Report: There are no signs of hyperdensity, large infarct or hyperdense artery. The infarct extent is under 33%. Patient suitable for thrombolysis.

Figure 2.5. A patient's case in the database (example for illustration purposes only): fourteen slices and a short radiological report.

[1] Used in the NeuRadIR retrieval system at http://dcs.vein.hu/CIR.

The cases were indexed using relevant medical terms (figure 2.6.a) in the written reports and a set of criteria relative to image content (figure 2.6.b). The database contains 68 index terms. A controlled vocabulary was created based on both textual reports and standard specialist queries.

- Aphasia
  - sudden
  - global
- Bedridden
- Cardiac arrhythmia
- Collapse
  - Sudden
- Coma
- Consciousness
  - undisturbed
  - impaired

- Coronary artery stenosis
- Eye deviation
  - conjugated
- Palsy gaze
- Myocardial infarction
- Orientation
  - undisturbed
  - impaired
  - completely disturbed
- Somnolence
- Stupor

- Hemiparesis
  - progressive
  - severe
  - sudden
  - slight
    - moderate
    - very severe
    - global
    - facial
    - subacute
    - left-sided
    - right-sided

Figure 2.6. a) Examples of relevant medical terms in written reports used as index terms. For example, the term ‘palsy gaze’ has Boolean values (Yes, No), whilst the term ‘moderate hemiparesis’ has the weight 0.5 in the original document-term matrix.

Hyperdensity,
Haemorrhage,
Infarction,
Hyperdense artery,
Hypodensity,
Thrombolysis,
Tissue volume,
Deformation,

Figure 2.6. b) Examples of criteria expressing relevant image features.

Both the medical terms and criteria were assigned weights. Thus, a *term-by-document* matrix *D* was constructed, where $d_{i,j}$ denoted the numeric value assigned to term (or criteria) *i* for case *j* (corresponding to a ‘document’ D_j). Table 2.1 shows a partial *term-by- document* matrix — with ten documents (from document *7b* to document *12a*) and ten index terms (from t_{28} to t_{37}) — of the medical database using *term frequency* weighting scheme.

Table 2.1. Partial *term-by-document* matrix of the medical database

	$\mathbf{D_{7b}}$	$\mathbf{D_{8a}}$	$\mathbf{D_{8b}}$	$\mathbf{D_{9a}}$	$\mathbf{D_{9b}}$	$\mathbf{D_{10a}}$	$\mathbf{D_{10b}}$	$\mathbf{D_{11a}}$	$\mathbf{D_{11b}}$	$\mathbf{D_{12a}}$
$\mathbf{w_{28}}$	1	0	0	0	1	0	0	0	1	0
$\mathbf{w_{29}}$	1	1	1	1	1	1	1	1	1	1
$\mathbf{w_{30}}$	0	0	0	0	0	0	0	0	0	0
$\mathbf{w_{31}}$	1	1	0	1	1	1	1	0	0	1
$\mathbf{w_{32}}$	0	0	0	0	0	1	0	0	0	0
$\mathbf{w_{33}}$	0	0	0	0	0	0	0	0	0	0
$\mathbf{w_{34}}$	1	1	0	1	0	1	1	1	0	1
$\mathbf{w_{35}}$	1	0	0	0	0	0	1	0	0	0
$\mathbf{w_{36}}$	0	0	0	1	0	0	0	0	0	0
$\mathbf{w_{37}}$	1	1	1	1	1	1	1	1	1	0

CHAPTER 3

HYPERBOLIC INFORMATION RETRIEVAL

In this chapter the HIR (Hyperbolic Information retrieval) Model is introduced, where the similarity measure is derived from the Cayley-Klein hyperbolic distance. It is shown formally, as well as experimentally that the HIR model is equivalent with the Cosine-based Vector Space Model using normalised weighting scheme. The first application using the HIR model — called NeuRadIR — is also presented.

3.1 Motivation

One of the most widely used models of information retrieval to process texts efficiently and retrieve information is the Vector Space Model (VSM). The Euclidean geometry has been the only type of space used in the VSM in general, but non-Euclidean geometry is becoming increasingly important in modern science and technology. The application of non-Euclidean spaces to information processing in general seems to experience its beginnings: they are used for information visualisation. In a hyperbolic space the area of a circle grows exponentially with respect to its radius, whereas in Euclidean space the area only grows quadratically. Thanks to this property a convenient way to visualize exponentially growing trees can be derived (Phillips and Gunn, 1992; Phillips, Levy and Munzner, 1993). They draw 3D hyperbolic pictures of large hierarchies or graphs (such as the Web) in the interior of a ball, use Euclidean straight lines, but the way distance is measured is changed. Thus, an effective way to visualise structures is obtained (more can be represented in less space, although in a distorted way; in a "fisheye" view style).

This chapter investigates the possibility of defining a VSM in a hyperbolic space. A non-Euclidean space is applied to IR by defining a VSM in the hyperbolic space with a hyperbolic similarity measure. It is shown that the new model (HIR = Hyperbolic Information Retrieval) is equivalent with the Cosine-based VSM with normalised weighting scheme.

3.2 Vector Space Model

All theoretical and practical research in IR is based on a few basic models which have been elaborated over time. Depending on how the documents, query and retrieval are modelled different formal methods can be distinguished in IR. These models are based on ideas and techniques form different scientific fields such as mathematics, logics, information science, artificial intelligence and quantum theory (Dominich, 2002).

The first models of IR are based on mathematical techniques because using mathematical knowledge was well known and understood, so these models could be created easily. In these models, retrieval of information is based on the mathematical concept "distance" (or similarity) between the query and objects. Different specific versions of these models are used in commercial retrieval system, so these models are considered to be classical models of IR, namely Boolean, Vector Space and Probabilistic Model. Models applying logics and information science are relatively new, where the retrieval of information is based on some inference processes or flow of information between the query and the objects to be searched. These models are called non-classsical models (Dominich, 2002): Information Logic, Situation Theory and Interaction Information retrieval Model. The models of IR applying different AI (Artificial Intelligence) or AI-related methods are called alternative models. They enhance the classical models of IR. In models that are based on ideas and principles from Quantum Mechanics, retrieval of information is a result of an effective and real interaction between the query and the objects to be searched. The alternative IR models are: Cluster Model, Fuzzy Model, Latent Semantic Indexing Model, Alternative Probabilistic Model (Inference Network Model (Turtle and Croft, 1990, 1991), Belief Network Model (Ribeiro-Neto and Muntz, 1996)) and Artificial Intelligence Based Model.

In the field of information retrieval, the Vector Space Model (VSM) is an important, well-understood and extensively researched classical model, which has been widely used to process texts efficiently and retrieve information for some forty years (Salton, 1966). It is called VSM because each document and query is mapped to a point in the feature space based on frequencies of keywords appearing in the text. The feature space is mathematically modelled by the orthonormal Euclidean space, i.e., the space (or geometry) is defined by a system of pairwise perpendicular coordinate axes corresponding to index terms. Retrieval is based on whether the "query vector" and the "document vector" are close enough.

Given a finite set D of elements called *documents*:

D_j, , $j = 1, \dots, m \in$ **N** (**N** denotes the set of natural numbers),

and a finite set T of elements called index *terms*:

t_i, $i = 1, \dots, n \in$ **N** (**N** denotes the set of natural numbers).

In the *Vector Space Model* (van Rijsbergen, 1979; Salton and McGill, 1983; Baeza-Yates and Ribeiro-Neto, 1999), briefly VSM, of information retrieval, every document D_j is assigned a vector:

$\mathbf{w}_j = (w_{ij})_{i=1,\dots,n}$ of *weights*, where $w_{ij} \in \mathbf{R}$ ($\mathbf{R}$ denotes the set of real numbers) denotes the *weight* of term t_i for document D_j.

The matrix $W = (w_{ij})_{n\times m}$ is called the *term-by-document matrix*.

3.2.1 Weighting Schemes

The general form of a *weighting scheme* (Berry and Browne, 2000) is as follows:

$$w_{ij} = local\_weight_{ij} \times global\_weight_i \times normalisation_j = l_{ij} \times g_i \times n_j$$

The types of *local term weights* l_{ij} are as follows:

- *b* (Binary): $\chi(f_{ij})$
- *l* (Logarithmic): $\log(1+f_{ij})$
- *t* (Term frequency): f_{ij}
- *n* (Augmented normalized term frequency): $(\chi(f_{ij})+(f_{ij}/\max_k f_{kj}))/2$

Generally used formulas for *global term weights* g_i are as follows:

- *x* (None): 1
- *e* (Entropy): $1+\left(\sum_{j=1}^{m}\left(p_{ij}\cdot\log p_{ij}\right)/\log m\right)$
- *p* (Probabilistic Inverse) $\log\left(\left(m-\sum_{j=1}^{m}c\left(f_{ij}\right)\right)/\sum_{j=1}^{m}c\left(f_{ij}\right)\right)$
- *f* (Inverse document frequency, or IDF): $\log\left(\frac{m}{\sum_{j=1}^{m}f_{ij}}\right)$
- *n* (Normal): $\frac{1}{\sqrt{\sum_{j=1}^{m}f_{ij}^{2}}}$

Formulas for document *normalisation* n_j, are as follows:

- *x* (None): 1

- *c* (Cosine): $\left(\sum_{i=1}^{n} (g_i \cdot l_{ij})^2 \right)^{-1/2}$

The choice for:

(i) the local weight l_{ij} depends on the vocabulary (e.g., technical, scientific, journal articles, magazines, encyclopaedia) or words used,

(ii) the global weight g_i depends on the change rate (static, often changing) of the collection,

(iii) the *normalisation* n_j depends on document length.

Widely used weighting schemes are as follows:

- *f* (term frequency): $w_{ij} = l_{ij} = f_{ij}$

- *maxNorm* : $w_{ij} = l_{ij} = \dfrac{f_{ij}}{\max_k \; f_{kj}}$

- *tf-idf* (term_frequency × IDF): $w_{ij} = l_{ij} \times g_i = f_{ij} \times \log \dfrac{m}{F_i}$

- *n-idf* (normalised IDF): $w_{ij} = \dfrac{f_{ij} \cdot \log \dfrac{m}{F_i}}{\sqrt{\sum_{i=1}^{n} \left(f_{ij} \cdot \log \dfrac{m}{F_i} \right)^2}}$

- *tfn* (term frequency × normalised): $w_{ij} = l_{ij} \times g_i = \dfrac{f_{ij}}{\sqrt{\sum_{i=1}^{n} f_{ij}^2}}$

where, f_{ij} denotes the number of occurrences of term t_i in document D_j,

F_i is the number of documents in which the term t_i occurs.

For convenience $c(r) = \begin{cases} 1, r > 0 \\ 0, r < 0 \end{cases}$, probability $p_{ij} = f_{ij} / \sum_{j=1}^{m} f_{ij}$.

For technical or scientific vocabularies schemes of the form *xnx,* with normalised term frequencies are generally recommended. For more general vocabularies simple term frequencies (t**) may be sufficient. When the term list is relatively short, such in case of controlled vocabulary, binary term frequencies (b**) are useful (Berry and Browne, 2000).

3.2.2 Similarity Measures

Let Q denote a user's *query* and

$\mathbf{q} = (q_i)_{i=1,\dots,n}$ the corresponding query weight vector.

The vectors $\mathbf{w}_j$ and $\mathbf{q}$ belong to the E_n Euclidean orthonormal space, in which the weights $\mathbf{w}_j$ and $\mathbf{q}$ are regarded as Cartesian coordinates (of points corresponding to document D_j and query Q). In other words, each term t_i is assigned to an axis x_i, all the axes intersect each other in one common point O (called the origin), they are pairwise perpendicular to each other in the origin, and the weight w_{ij} corresponds to a point on the axis x_i (one separate point for each document D_j). Thus, every document D_j $(j = 1, \dots, m)$ is represented by a vector w_j, which defines a point in the space E_n.

The *relevance* of document D_j relative to query Q is given by the value of a *similarity measure* $\sigma(\mathbf{w}_j, \mathbf{q})$, whose general form is as follows:

$$\sigma(\mathbf{w}_j, \mathbf{q}) = \frac{\mathbf{w}_j \mathbf{q}}{\Delta}$$

where $\mathbf{w}_j\mathbf{q}$ denotes the inner — or dot — product of the vectors $\mathbf{w}_j$ and $\mathbf{q}$.

A function $\sigma(\mathrm{w}, \mathbf{q})$ is similarity if it satisfies the three similarity properties (Van Rijsbergen, 1979), i.e.,

$\sigma: D \times D \rightarrow \mathbf{R}$

– normalisation: $0 \le \sigma(\mathbf{w}, \mathbf{q}) \le 1$,

– symmetry: $\sigma(\mathbf{w}, \mathbf{q}) = \sigma(\mathbf{q}, \mathbf{w})$, $\forall \mathbf{q}, \mathbf{w}$; i.e., the order in which the query and the document are considered when computing the similarity value is indifferent;

– reflexivity: $\mathbf{w} = \mathbf{q} \Rightarrow \sigma(\mathbf{w}, \mathbf{q}) = \kappa$; i.e., the value of the similarity measure is equal to a predefined and fixed maximal value κ if the query and the document are exactly the same; the reverse is not necessarily true; for example, if σ is normalised then κ may be taken as being equal to 1.

Depending on the formula used to calculate the denominator Δ several well-known similarity measures have been proposed over time such as:

Dot product: $$\sigma\,(\mathbf{w}_j,\,\mathbf{q}) = \sum_{i=1}^{n} w_{ij}q_i$$

Cosine measure: $$\sigma\,(\mathbf{w}_j,\,\mathbf{q}) = \frac{\sum_{i=1}^{n} w_{ij}q_i}{\sqrt{\sum_{i=1}^{n} w_{ij}^2 \cdot \sum_{i=1}^{n} q_i^2}}$$

Dice's coefficient: $$\sigma\,(\mathbf{w}_j,\,\mathbf{q}) = \frac{\sum_{i=1}^{n} w_{ij}q_i}{\sum_{i=1}^{n} (w_{ij} + q_i)}$$

Jaccard's coefficient: $$\sigma\,(\mathbf{w}_j,\,\mathbf{q}) = \frac{\sum_{i=1}^{n} w_{ij}q_i}{\sum_{i=1}^{n} \left(\frac{w_{ij} + q_i}{2^{w_{ij}q_i}} \right)}$$

Overlap coefficient: $$\sigma\,(\mathbf{w}_j,\,\mathbf{q}) = \frac{\sum_{i=1}^{n} w_{ij}q_i}{\min\left(\sum_{i=1}^{n} w_{ij}, \sum_{i=1}^{n} q_i \right)}$$

In what follows, the Cosine measure will be used; namely, its explicit formula is $\Delta = \|\mathbf{w}_j\| \cdot \|\mathbf{q}\|$ ($\|.\|$ denotes the Euclidean norm of a vector).

EXAMPLE 3.1 (based on Berry and Browne, 2000)

Given a small collection of book titles (7 documents D_j) with 9 index terms $t_{i,}$ and a query Q as illustrated in table 3.1:

Table 3.1. Collection of book titles with the index terms

Terms		Documents		Query
T1	Baby	D1	Infant and Toddler First Aid	
T2	Child	D2	Babies and Children's Room (For your Home)	Child
T3	Guide	D3	Child Safety at Home	
T4	Health	D4	Your Baby's Health and Safety: From Infant to Toddler	
T5	Home	D5	Baby Proofing Basics	Home
T6	Infant	D6	Your Guide to Easy Rust Proofing	Infant
T7	Proofing	D7	Beanie Babies Collectors Guide	Proofing
T8	Safety			Safety
T9	Toddler			

Using *tfn* weighting scheme, the *term-by-document* matrix, and the *term-by-query* are the following:

$$D := \begin{pmatrix} 0 & 0.5774 & 0 & 0.4472 & 0.7071 & 0 & 0.7071 \\ 0 & 0.5774 & 0.5774 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.7071 & 0.7071 \\ 0 & 0 & 0 & 0.4472 & 0 & 0 & 0 \\ 0 & 0.5774 & 0.5774 & 0 & 0 & 0 & 0 \\ 0.7071 & 0 & 0 & 0.4472 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.7071 & 0.7071 & 0 \\ 0 & 0 & 0.5774 & 0.4472 & 0 & 0 & 0 \\ 0.7071 & 0 & 0 & 0.4472 & 0 & 0 & 0 \end{pmatrix} \qquad Q := \begin{pmatrix} 0 \\ 0.4472 \\ 0 \\ 0 \\ 0.4472 \\ 0.4472 \\ 0.4472 \\ 0.4472 \\ 0 \end{pmatrix}$$

Similarity values for the given Q using Cosine, Dice and Jaccard similarity measures are illustrated in table 3.2. It shows, that document D_3 is the most relevant to the user's query.

Table 3.2. Similarity values of VSM model using Cosine, Dice or Jaccard measures

Document	Similarity measure		
	Cosine	Jaccard	Dice
D1	0.316	0.092	0.087
D2	0.516	0.142	0.13
D3	0.775	0.224	0.195
D4	0.4	0.094	0.089
D5	0.316	0.092	0.087
D6	0.316	0.092	0.087
D7	0	0	0

3.2.3 Rank Order Preservation

Given any two documents (objects) D_1 and D_2, and any two similarity measures σ_1 and σ_2. If the two documents (objects) are ranked in the same order by these measures relative to any query Q, i. e.,

$$\sigma_1\,(\mathbf{w}_1, \mathbf{q}) \le \sigma_1\,(\mathbf{w}_2, \mathbf{q}) \Leftrightarrow \sigma_2\,(\mathbf{w}_1, \mathbf{q}) \le \sigma_2\,(\mathbf{w}_2, \mathbf{q}),\ \forall\, D_1, D_2, Q,$$

then the similarity measures σ_1 and σ_2 are said to preserve the rank order. The importance of rank order preservation consists in that all rank order preserving similarity measures are equivalent with each other. In other words, any of them can replace the others, or equivalently, they all can return the same documents (objects).

In the VSM, in general, the similarity measures do not preserve the rank order of the retrieved documents. Only practice and experimentation but no sound theoretical argument can recommend which one to use in order to obtain better (more relevant) results.

EXAMPLE 3.2

Given the following *term-by-document* matrix and the *term-by-query*:

$$D := \begin{pmatrix} 0.377 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0.377 & 0.377 & 0 & 0.5774 & 0 & 0.7071 & 0 & 0 & 0 & 0 \\ 0.377 & 0.377 & 0 & 0 & 0.5774 & 0.7071 & 0 & 0 & 0 & 0 \\ 0 & 0.377 & 0 & 0 & 0.5774 & 0 & 0 & 0 & 0 & 0 \\ 0.377 & 0.377 & 0 & 0.5774 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0.377 & 0.377 & 0.7071 & 0.5774 & 0.5774 & 0 & 0 & 0 & 0 & 1 \\ 0.377 & 0.377 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.377 & 0.7071 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.377 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \qquad Q := \begin{pmatrix} 0 \\ 0.4472 \\ 0 \\ 0 \\ 0.4472 \\ 0.4472 \\ 0.4472 \\ 0.4472 \\ 0 \end{pmatrix}$$

Similarity values for the given Q using Cosine, Dice and Jaccard similarity measures are illustrated in table 3.3. Figure 3.1 shows these values in graphical form. It can be seen clearly, that these measures do not preserve the rank order (e.g. $Cosine_1$ > $Cosine_2$ but $Jaccard_2$ > $Jaccard_1$, where $Cosine_i$ denotes the similarity value of the document D_i using Cosine measure, and respectively $Jaccard_i$ denotes the similarity value of the document D_i using Jaccard measure).

Table 3.3. Similarity values of VSM model using Cosine, Dice or Jaccard measures

Document	**Similarity measure**		
	Cosine	**Jaccard**	**Dice**
D0	0.676	0.149	0.138
D1	0.845	0.191	0.173
D2	0.632	0.198	0.173
D3	0.775	0.224	0.195
D4	0.258	0.068	0.065
D5	0.316	0.092	0.087
D6	0	0	0
D7	0	0	0
D8	0.447	0.157	0.138
D9	0.447	0.157	0.138

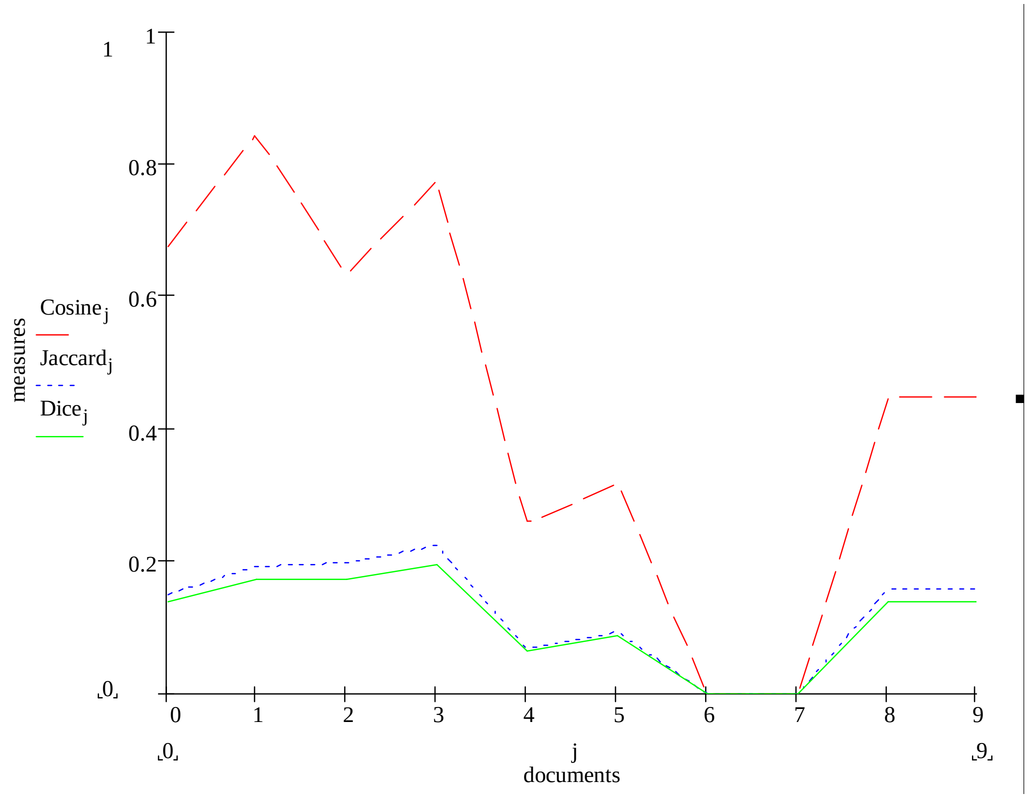


Figure 3.1. Visualization of rank order in Cosine, Jaccard and Dice measures using *tfn* weighting scheme.

It can be seen that, for example the Cosine measure and Jaccard coefficient do not preserve the rank order.

3.2.4 The s-Space

All the similarity measures may be viewed as normalised versions of the Dot product, which is a measure of how many index terms the query Q and document D_j have in common, and to what extent; i.e., a measure of how many times $q_i \neq w_{ji}$. All the similarity measures are normalised (Cosine, Dice, Jaccard).

The concept of a σ-*space*, introduced in (Dominich, 2001), is a formal generalisation of the VSM in order to emphasize the fact that retrieval is based on similarity measures. A set D of objects with a symmetric and reflexive similarity measure, i.e.,

$$\sigma: D \times D \to \mathbf{R}$$

- symmetry: $\sigma(a, b) = \sigma(b, a)$, $\forall a, b \in D$; i.e., the order in which the query and the document are considered when computing the similarity value is indifferent;
- reflexivity: $a = b \Rightarrow \sigma(a, b) = \kappa$; i.e., the value of the similarity measure is equal to a predefined and fixed maximal value κ if the query and the document are exactly the same; the reverse is not necessarily true; for example, if σ is normalised then κ may be taken as being equal to 1.

is referred to as a σ-*space*. It was shown that the following holds (Dominich, 2001), stated here without proof:

THEOREM 3.1 (Dominich, 2001) Let $\langle E, \mu \rangle$ be a (pseudo-) metric space (μ is normalised, which is always possible). Then,

(i) the induced topological space is a σ-space on E, and

(ii) $\langle E, 1 - \mu \rangle$ is a σ-space. ♦

The importance of Theorem 3.1 consists in that it allows for constructing a similarity measure from a given (pseudo-) metric, and this property will be used in my thesis.

3.3 Cayley-Klein Hyperbolic Geometry

In this section, the Cayley-Klein hyperbolic geometry is briefly described in a form used in my dissertation.

3.3.1 Non-Euclidean Geometry

Non-Euclidean geometry (Bolyai, 1987) is a geometry that is different from the Euclidean (Classical) geometry in that Euclid’s fifth postulate (in plane, there exists exactly one parallel line to a given line through a given point that is not on the given line) does not hold. One of the most useful non-Euclidean geometries is elliptic or spherical geometry, which describes the surface of a sphere. It replaces the parallel postulate with the statement "through any point in the plane, there exist no lines parallel to a given line".

The other non-Euclidean geometry is called hyperbolic (or Bolyai–Lobachevsky) geometry, which is a "curved" space. Hyperbolic geometry satisfies Euclid’s postulates except the fifth, i.e. for any hyperbolic line *l*, and point *p* not on *l*, there exist at least two hyperbolic lines through *p* and parallel to *l*.

3.3.2 Cayley-Klein Model

The Cayley-Klein hyperbolic geometry or space (C-KHS) — or model — is an example for hyperbolic geometry.

Let $\mathbf{R}^n$ denote the Euclidean (orthonormal) space (Császár, 1974).

Let *A* and *B* denote two points in $\mathbf{R}^n$, and $(x_1, x_2, ..., x_n)$ and $(y_1, y_2, ..., y_n)$ denote their Cartesian coordinates, respectively.

The Euclidean distance $d_E(AB)$ between the points *A* and *B* is defined as follows (Patterson and Rutherford, 1965):

$$d_E(AB) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{3.1}$$

Let

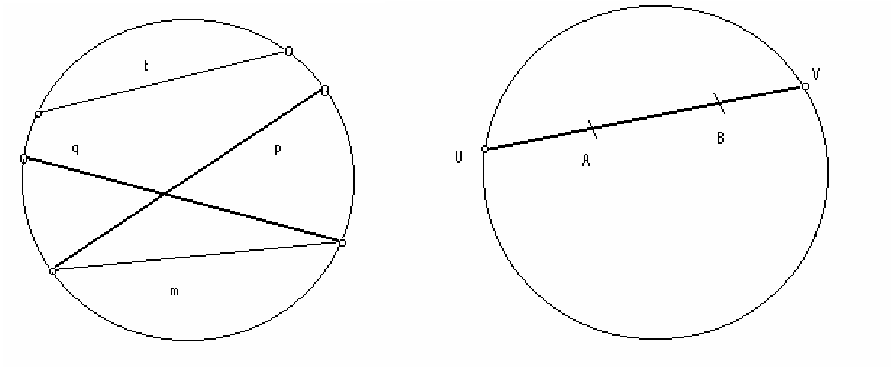
$$S(O,r) = \left\{ P(x_1, x_2, ..., x_n) \middle| \sum_{i=1}^{n} x_i^{\ 2} < r^2 \right\} \tag{3.2}$$

denote the interior of a hyper-sphere *S* having its centre in the origin *O* of the space $\mathbf{R}^n$, and radius $r \in \mathbf{R}$, $r > 0$.

The *S*(*O*, *r*) hyper-sphere is the C-KHS *space*.

The *points P* of the C-KHS are all the points of *S*(*O*, *r*), i.e, $P \in S(O, r)$.

The *lines* of the C-KHS space are open chords of the hyper-sphere *S* deprived of their endpoints. If the lines *m* and *q* have a common endpoint (on the boundary of the hyper-sphere in the Euclidean space) they are referred to as *asymptotically parallel* (figure 3.2.a). The scientific role and importance of non-Euclidean geometries is well-known (Anderson, 1999). In my context, however, the hyperbolic distance rather than parallelism will play an important role.



(a) (b)

Figure 3.2. a) m, p, q, t are lines in the C-KHS. Notice that the endpoints do not belong to the C-KHS, nor do the points of the circle (the circle is only drawn to show the 'limits' of the hyperbolic space). The lines p and q are asymptotically parallel to line m: m || p, m || q. The line t is divergently parallel to line m: m || t.

b) Example for a line segment AB in the C-KHS.

Note: The C-KHS space satisfies Hilbert's axioms on incidence, ordering and congruence, as well as Archimedes' and Cantor's axioms on continuity (Hilbert and Cohn-Vossen, 1932), and is thus a continuous absolute space.

The concept of a distance is defined using that of a cross ratio.

The hyperbolic length $d_H(AB)$of the line segment *AB* is defined as the cross-ratio of the points *U, A, B, V* (figure 3.2.b) as follows:

$$d_H(AB) = k \cdot \left| \ln \frac{d_E(AU) \cdot d_E(BV)}{d_E(AV) \cdot d_E(BU)} \right| \tag{3.3}$$

where $k \in \mathbf{R}$ is a positive constant,

U and V are the points of intersection of the Euclidean line through the points A and B with the hyper-sphere.

In the following, it will be assumed, without loss of generality, that *k* = 1.

The hyperbolic distance satisfies the properties of the metric:

- Non-negativity: the hyperbolic distance is non-negative (immediate from its definition):

$$d_H(AB) \geq 0, \forall A, B \in \text{C-KHS} \tag{3.4}$$

- Symmetry: the hyperbolic distance is symmetric:

$$d_H(AB) = \left| \ln \frac{d_E(AU)\cdot d_E(BV)}{d_E(AV)\cdot d_E(BU)} \right| = \left| \ln \frac{1}{\dfrac{d_E(AV)\cdot d_E(BU)}{d_E(AU)\cdot d_E(BV)}} \right| =$$

$$= \left| \ln 1 - \ln \frac{d_E(AV)\cdot d_E(BU)}{d_E(AU)\cdot d_E(BV)} \right| = \left| \ln \frac{d_E(AV)\cdot d_E(BU)}{d_E(AU)\cdot d_E(BV)} \right| = d_H(BA)$$

$\forall\, A, B \in$ C-KHS (3.5)

- Reflexivity: the hyperbolic distance is reflexive:

$A = B \Rightarrow d_H(AB) = d_H(AA) \Rightarrow d_E(BU) = d_E(AU)$

and $d_E(BV) = d_E(AV)$

$\Rightarrow |\ln 1| = 0$ (3.6)

- Triangle inequality: the hyperbolic distance satisfies the triangle inequality:

$d_H(AB) + d_H(BC) =$

$$= \left| \ln \frac{d_E(AU)\cdot d_E(BV)}{d_E(AV)\cdot d_E(BU)} \right| + \left| \ln \frac{d_E(BU)\cdot d_E(CV)}{d_E(BV)\cdot d_E(CU)} \right| \geq$$

$$\geq \left| \ln \frac{d_E(AU)\cdot d_E(BV)}{d_E(AV)\cdot d_E(BU)} + \ln \frac{d_E(BU)\cdot d_E(CV)}{d_E(BV)\cdot d_E(CU)} \right| =$$

$$= \left| \ln \frac{d_E(AU)\cdot d_E(BV)}{d_E(AV)\cdot d_E(BU)} \times \frac{d_E(BU)\cdot d_E(CV)}{d_E(BV)\cdot d_E(CU)} \right| =$$

$$= \left| \ln \frac{d_E(AU)\cdot d_E(CV)}{d_E(AV)\cdot d_E(CU)} \right| = d_H(AC),$$

$\forall\, A, B, C \in$ C-KHS (3.7)

From the hyperbolic distance — because it satisfies the metrics properties — a similarity measure can be derived using Theorem 3.1.

3.4 Hyperbolic Information Retrieval (HIR) Model [THESIS 1.a]

The Hyperbolic Information Retrieval (HIR) Model is proposed in [P1, P6] using a similarity measure derived from the hyperbolic distance.

Given a VSM.

Let **R'**^{n} denote the n-dimensional Euclidean space obtained by translating the space **R**^{n} into the query-point Q, i.e., the origin O of **R**^{n} is translated into Q.

Let us consider the following C-KHS:

$$S'(O',r) = S'(Q(q_1,q_2,...,q_{i,},...,q_n),r) = \left\{ D(d_1,d_2,...,d_n) \middle| \sum_{i=1}^{n}(d_i - q_i)^2 < r^2 \right\} \quad (3.8)$$

So the radius r of the C-KHS must be the following; because it is required to use all the documents for retrieval:

$$r > \max_{D} d_E(QD)$$

Using the definition of the hyperbolic distance, the hyperbolic distance $d_H(QD)$ in the translated space $S'(Q, r)$ is as follows:

$$d_H(QD) = \left| \ln \frac{d_E(QU) \cdot d_E(DV)}{d_E(QV) \cdot d_E(DU)} \right| = \left| \ln \frac{r \cdot (r - d_E(QD))}{r \cdot (r + d_E(QD))} \right| =$$

$$= \left| \ln \frac{r - \sqrt{\sum_{i=1}^{n}(w_{ij} - q_i)^2}}{r + \sqrt{\sum_{i=1}^{n}(w_{ij} - q_i)^2}} \right| \quad (3.9)$$

Based on Theorem 3.1, the C-KHS space $S'(Q, r)$ can be turned into a σ-space as follows:

(i) the hyperbolic distance d_H is normalised, for example, by taking $d_H = d_H / (1 + d_H)$; this is required for the similarity measure to be positive and smaller than unity;

(ii) a function s_H is defined as follows:

$s_H(\mathbf{w}, \mathbf{q}) = 1 - d_H(QD)$

Thus, the explicit form of the Hyperbolic similarity measure $s_H(\mathbf{w}, \mathbf{q})$ is as follows:

$$s_H(w,q) = \frac{1}{1 + \left| \ln \dfrac{r - \sqrt{\sum_{i=1}^{n} \left(w_{ij} - q_i\right)^2}}{r + \sqrt{\sum_{i=1}^{n} \left(w_{ij} - q_i\right)^2}} \right|} = \frac{1}{\ln \dfrac{e}{\dfrac{r - \sqrt{\sum_{i=1}^{n} \left(w_{ij} - q_i\right)^2}}{r + \sqrt{\sum_{i=1}^{n} \left(w_{ij} - q_i\right)^2}}}} =$$

$$s_H(w,q) = \left(\ln \left(e \cdot \frac{r + \sqrt{\sum_{i=1}^{n} \left(w_{ij} - q_i\right)^2}}{r - \sqrt{\sum_{i=1}^{n} \left(w_{ij} - q_i\right)^2}} \right) \right)^{-1} \qquad (3.10)$$

It is shown that:

THEOREM 3.2 The function s_H is a similarity measure.

Proof:

Based on (Van Rijsbergen, 1979), the function $s_H(\mathbf{w}, \mathbf{q})$ is a similarity measure if it satisfies the three similarity properties:

- Normalisation: $0 \le \sigma(w, q) \le 1$,

$$0 \le \left(\ln \left(e \cdot \frac{r + \sqrt{\sum_{i=1}^{n} \left(w_{ij} - q_i\right)^2}}{r - \sqrt{\sum_{i=1}^{n} \left(w_{ij} - q_i\right)^2}} \right) \right)^{-1} \le 1 \qquad (3.11)$$

- Symmetry: $\sigma(a, b) = \sigma(b, a), \forall a, b \in D$

$$d_E(QD) = d_E(DQ) \text{ Þ } s_H(Q, D) = s_H(D, Q) \qquad (3.12)$$

- Reflexivity: $a = b \Rightarrow \sigma(a, b) = \kappa, \forall a, b \in D$

$$Q = D \Rightarrow d_E(QD) = d_E(QQ) = 0 \Rightarrow s_H(Q, Q) =$$

$$(\ln(e \cdot \frac{r}{r}))^{-1} = (\ln e)^{-1} = 1 \qquad (3.13)$$

Hence, s_H is a similarity measure.♦

An additional property of the hyperbolic distance used in chapter 4.6 is as follows :

d_H becomes infinitely large when either (point B)of the points approaches the surface (point V) of the hyper-sphere, i.e.,

$$B \circledR V \Rightarrow \; d_E(BV) = 0 \Rightarrow \lim_{B \to V} d_H(AB) = +\infty \tag{3.14}$$

HIR model based on Cayley-Klein hyperbolic geometry was introduced in this chapter. The measure s_H was derived from the hyperbolic distance, and it was proved that it satisfies the properties of the similarity measure.

Hereafter, let denote *Hyp*(**w**, **q**) the similarity measure: s_H(**w**, **q**).

3.5 VSM and HIR: Equivalent Models [THESIS 1.b]

It is shown (formally, and experimentally) in [P1, P5] that the HIR model and the VSM equipped with the Cosine measure are equivalent in an important practical case.

3.5.1 Equivalence of VSM and HIR: Formal Proof

Two IR models or systems are equivalent if they produce the same ranking (Dominich, 2001). As known (e.g., Meadow, Boyce and Kraft, 1999; Berry and Browne, 2000), for technical disciplines, the usage of the *tfn* (normalised term frequency) weighting scheme is recommended as yielding good results. It is shown that the Cosine similarity measure and the Hyperbolic measure preserve the rank order of documents under this weighting scheme.

THEOREM 3.3: The Cosine and Hyperbolic similarity measures preserve the rank order under the *tfn* weighting scheme.

Proof:

Let $f_{i,j}$ mean the number of occurrences of term t_i in document D_j. (f_{ij} could be any other value, in fact.) The normalised weighting scheme means that the terms are assigned normalised weights w_{ij} as follows:

$$w_{ij} = \frac{f_{ij}}{\sqrt{\sum_{i=1}^{n} f_{ij}^2}} \tag{3.15}$$

The query terms are assigned weights similarly.

Because $\sum_{i=1}^{n} w_{ij}^2$ = 1, $\sum_{i=1}^{n} q_i^2$ = 1 under this weighting scheme, the Cosine and Hyperbolic measures Hyp_j become:

$$Cosine_j = \frac{\sum_{i=1}^{n} w_{ij} q_i}{\sqrt{\sum_{i=1}^{n} w_{ij}^2 \cdot \sum_{i=1}^{n} q_i^2}} = \sum_{i=1}^{n} w_{ij} q_i \tag{3.16}$$

$$Hyp_j = \left(\ln\left(e \cdot \frac{r + \sqrt{\sum_{i=1}^{n} \left(w_{ij} - q_i\right)^2}}{r - \sqrt{\sum_{i=1}^{n} \left(w_{ij} - q_i\right)^2}} \right) \right)^{-1} = \left(\ln\left(e \cdot \frac{r + \sqrt{2 - 2 \cdot \sum_{i=1}^{n} w_{ij} \cdot q_i}}{r - \sqrt{2 - 2 \cdot \sum_{i=1}^{n} w_{ij} \cdot q_i}} \right) \right)^{-1} \tag{3.17}$$

respectively.

Thus, one can write the following equivalence:

$$\sum_{i=1}^{n} w_{ij} \cdot q_i \le \sum_{i=1}^{n} w_{ik} \cdot q_i \;\Leftrightarrow\; \sqrt{2 - 2 \cdot \sum_{i=1}^{n} w_{ij} \cdot q_i} \ge \sqrt{2 - 2 \cdot \sum_{i=1}^{n} w_{ik} \cdot q_i} \tag{3.18}$$

and using the property of *r* radius of C-KHS, that:

$$r > \max_{D} d_E(QD)$$

it follows:

$$Cosine_j \le Cosine_k \quad \Leftrightarrow \; Hyp_j \le Hyp_k \; \blacklozenge \tag{3.19}$$

In other words, a given VSM based on the Cosine measure and using the *tfn* weighting scheme can be replaced with a hyperbolic IR model (producing exactly the same answers and ranking).

3.5.2 Equivalence of VSM and HIR: Experimental Results

Experiment using a part of the medical database (section 2.2.4) is performed to illustrate of the rank order preservation in VSM and HIR.

Ten documents (from document 7b to document 12a) and ten index terms (from t_{28} to t_{37}) — with the connected weights (from w_{28} to w_{37}) — were used for the experiments. The query terms are: t_{28}, t_{30}, t_{31}, t_{36}, t_{37}. The ***term-by-document matrix*** (table 3.4), and the ***term-by-query vector*** (figure 3.3) using *tfn* weighting scheme were computed by Mathcad 2001i Professional Software.

Table 3.4. *Term-by-document* matrix using *tfn* weighting scheme.

	D_{7b}	D_{8a}	D_{8b}	D_{9a}	D_{9b}	D_{10a}	D_{10b}	D_{11a}	D_{11b}	D_{12a}
w_{28}	0.408	0.000	0.000	0.000	0.500	0.000	0.000	0.000	0.577	0.000
w_{29}	0.408	0.500	0.707	0.447	0.500	0.447	0.447	0.577	0.577	0.577
w_{30}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
w_{31}	0.408	0.500	0.000	0.447	0.500	0.447	0.447	0.000	0.000	0.577
w_{32}	0.000	0.000	0.000	0.000	0.000	0.447	0.000	0.000	0.000	0.000
w_{33}	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
w_{34}	0.408	0.500	0.000	0.447	0.000	0.447	0.447	0.577	0.000	0.577
w_{35}	0.408	0.000	0.000	0.000	0.000	0.000	0.447	0.000	0.000	0.000
w_{36}	0.000	0.000	0.000	0.447	0.000	0.000	0.000	0.000	0.000	0.000
w_{37}	0.408	0.500	0.707	0.447	0.500	0.447	0.447	0.577	0.577	0.000

$$q = \begin{pmatrix} 0.447 \\ 0 \\ 0.447 \\ 0.447 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.447 \\ 0.447 \end{pmatrix}$$

Figure 3.3. *Term-by-query* vector using *tfn* weighting scheme.

The computation of the cosine and hyperbolic similarity values for the query *q* was also performed using Mathcad. For the hyperbolic values the radius of the C-KHS was defined as 1.218 (r = 1.218). Table 3.5 shows the cosine and hyperbolic similarity values, while figure 3.4 illustrates these values in a graphical form. The solid line represents the similarity values of the Cosine measure and the dotted line shows the hyperbolic values. It can be seen clearly that for every two documents the similarity values follow the same order, i.e., the documents are ranked in exactly the same order. E.g. document D_{9b} is the most relevant to the query Q in the Cosine measure (0.671), and in the Hyperbolic measure (0.383) too.

Table 3.5. Cosine and hyperbolic (r = 1.218) similarity values

	D_{7b}	D_{8a}	D_{8b}	D_{9a}	D_{9b}	D_{10a}	D_{10b}	D_{11a}	D_{11b}	D_{12a}
Cosine	0.548	0.447	0.316	0.600	0.671	0.400	0.400	0.258	0.516	0.258
Hyp	0.323	0.277	0.204	0.348	0.383	0.254	0.254	0.029	0.309	0.029

The returned hit lists are the following:

$D_{9b,}$
$D_{9a,}$
$D_{7b,}$
$D_{11b,}$
$D_{8a,}$
$D_{10a,}$
$D_{10b,}$
$D_{8b,}$
$D_{11a,}$
D_{12a}

It can be seen clearly, that the rank is the same in both cases.

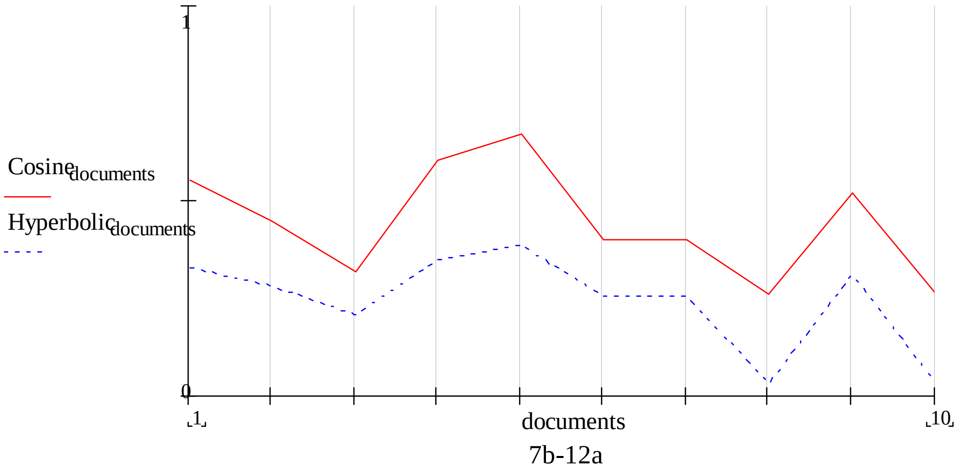


Figure 3.4. Visualisation of rank order preservation in VSM and HIR using *tfn* weighting scheme.

3.6 Application: NeuroRadiological Information Retrieval System

The application called NeuRadIR — in [P2, P4, P5, P7] — is a **Neu**ro**Rad**iological **I**nformation **R**etrieval System using the HIR model — besides two other retrieval techniques in order to satisfy different information need — to brain CT image and report retrieval. It was developed by our center named CIR (Center for Information retrieval) at the Department of Computer Science within the Cost Effective Health Preservation Consortium Project. The implemented system enables physicians (both radiologist and general practitioners) to use medical text and image database over the web in order to facilitate health preservation but also to assist diagnosis and patient care.

In medical practice most image retrieval systems are designed to help experienced physicians in diagnostic tasks and require that users have prior knowledge of the field and not capable for educational purpose (Guy, and Fftyche, 2000.).

The results of our research from the application viewpoint:

(i) Enhance the quality of specialist consultation as well as medical education.

(ii) General practitioner medical doctors may confirm a diagnosis or explore possible treatment plans through a consultation to the CT retrieval system over the Web.

(iii) Medical students may have images and would like to explore possible diagnoses or would like to see images corresponding to different pathological cases such as lesion, bleed or stroke.

3.6.1. System Description of NeuRadIR

The NeuRadIR application consists of computer program modules written in several languages as well as related documentation. The communication between the Web server and the search program is based on the CGI protocol. The *Report Editor* makes it possible to create/edit reports. The *CT Base Editor* makes it possible to create and modify the database containing the images and reports. The *Validation Module* consists of program, which carry out formal and consistency validation and statistics. The *Search Module* works online on the Web. Figure 3.5 illustrates the functional description of the retrieval system using UML use case diagram, and figure 3.6 shows the architecture of the implemented application.

Two databases — an English and a Hungarian one — were created for the NeuRadIR. English version was based on training material containing 40 cases. This medical database was described in section 2.2.4; it was used for some experiments in *Chapter 3,* and *4*. A controlled vocabulary was created based on both textual reports and standard specialist queries in both language.

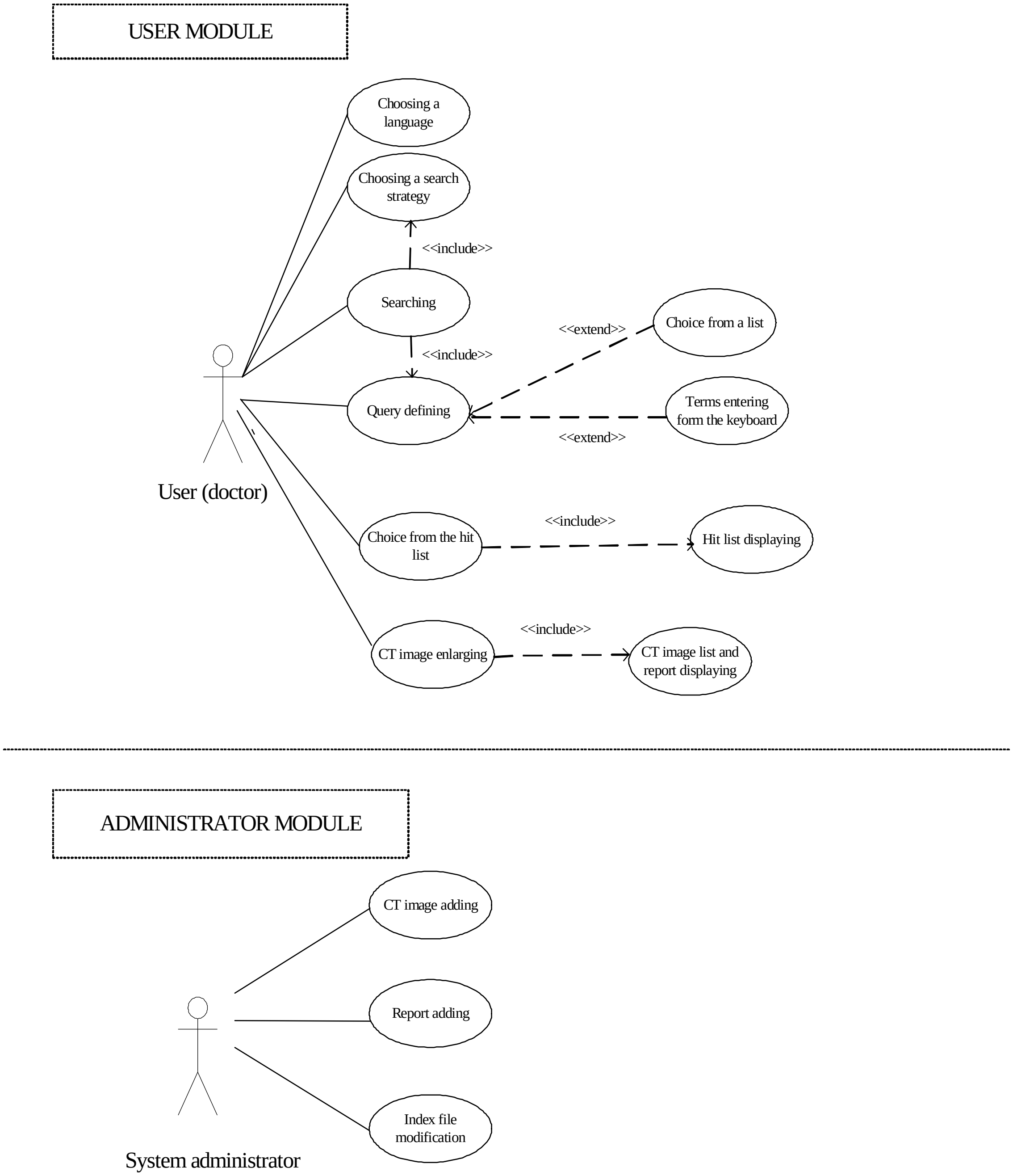


Figure 3.5. Functional description of the NeuRadIR using UML use case diagram.

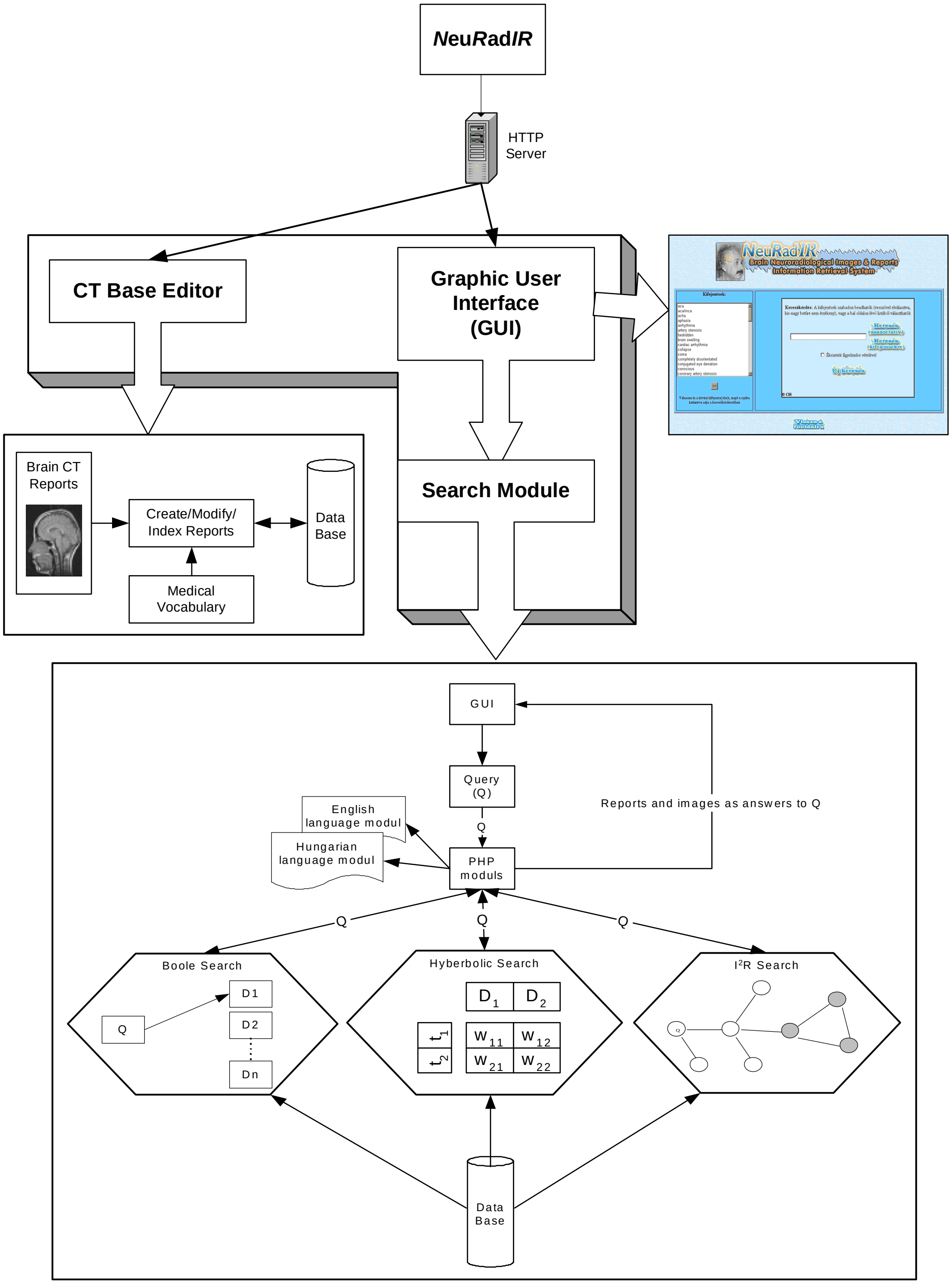


Figure 3.6. System architecture of the NeuRadIR

The search module is used online on the Web. Figure 3.7 shows "Search screen" of the application on which the user can enter the query. The query, i.e., medical terms (for example, brain) can be selected from the vocabulary or freely entered from the keyboard in the query line. By clicking on the SEARCH button the effective search is initiated in a local database on a server. Clicking on the BACK TO MAIN PAGE button takes us back to the title page. The user has two choices for selecting searching strategies by clicking on the "SEARCH associative" (using interaction information retrieval) or on the "SEARCH terms" (using HIR model) button.

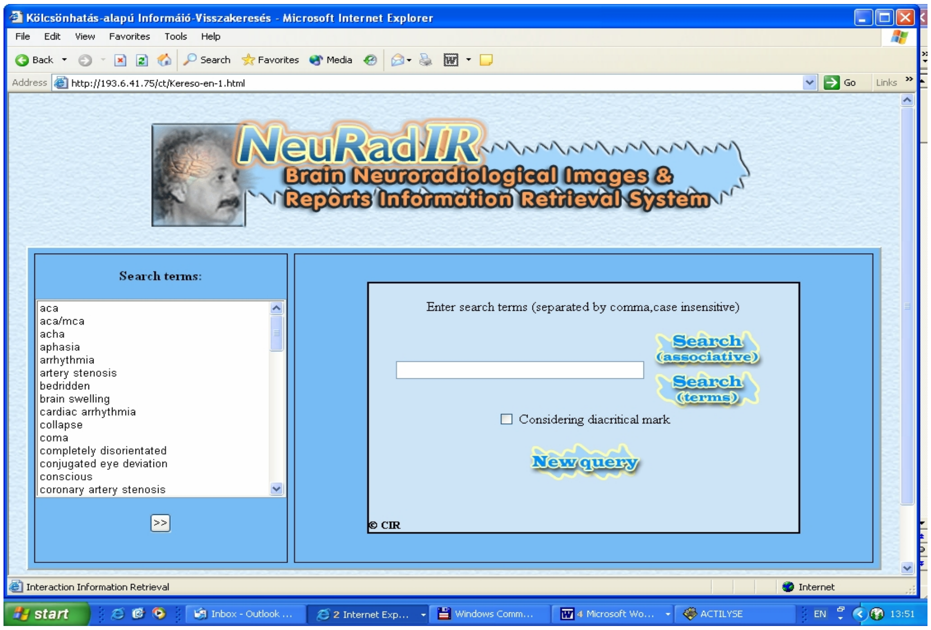
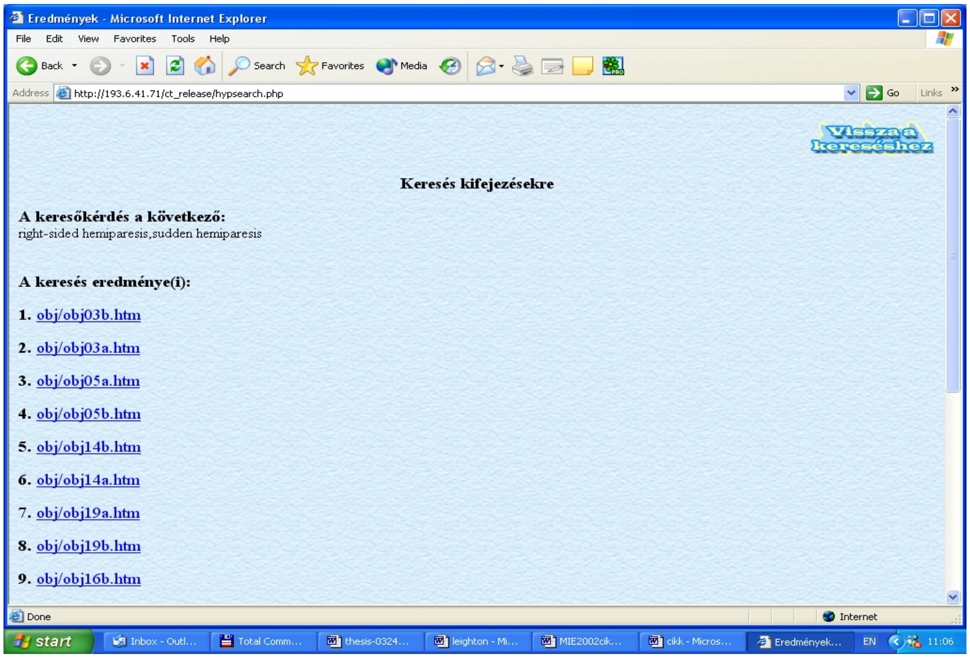

Kölcsönhatás-alapú Informáió-Visszakeresés - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://193.6.41.75/ct/Kereso-en-1.html
Go Links
NeuRadIR
Brain Neuroradiological Images & Reports Information Retrieval System
Search terms:
aca
aca/mca
acha
aphasia
arrhythmia
artery stenosis
bedridden
brain swelling
cardiac arrhythmia
collapse
coma
completely disorientated
conjugated eye deviation
conscious
coronary artery stenosis
>>
Enter search terms (separated by comma,case insensitive)
Search (associative)
Search (terms)
□ Considering diacritical mark
New query
© CIR
Interaction Information Retrieval
Internet
start
Inbox - Outlook ...
2 Internet Exp...
Windows Comm...
4 Microsoft Wo...
ACTILYSE
EN
13:51

Figure 3.7. Search screen of NeuRadIR

Figure 3.8 shows the returned hit list, and by clicking on any hit the textual information as well (figure 3.9) as the CT images (figure 3.10) are displayed.


Eredmények - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://193.6.41.71/ct_release/hypsearch.php
Go Links
Vissza a kereséshez
Done
Internet
start
Inbox - Outl...
Total Comm...
thesis-0324...
leighton - Mi...
MIE2002cik...
cikk - Micros...
Eredmények...
EN
11:06

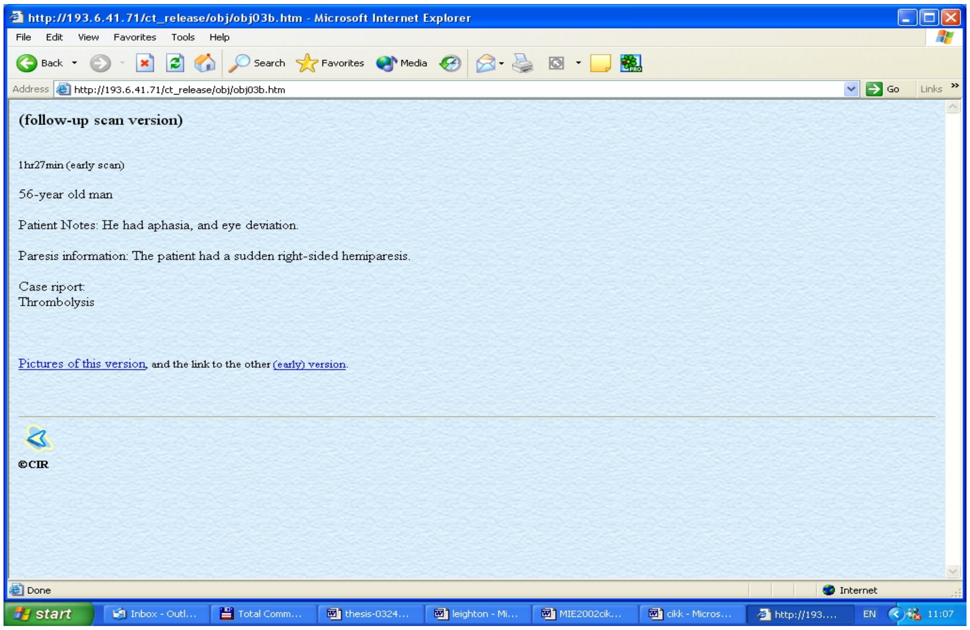
Keresés kifejezésekre

A keresőkérdés a következő:
right-sided hemiparesis,sudden hemiparesis

A keresés eredménye(i):

1. obj/obj03b.htm
2. obj/obj03a.htm
3. obj/obj05a.htm
4. obj/obj05b.htm
5. obj/obj14b.htm
6. obj/obj14a.htm
7. obj/obj19a.htm
8. obj/obj19b.htm
9. obj/obj16b.htm

Figure 3.8. Returned hit list of NeuRadIR.


http://193.6.41.71/ct_release/obj/obj03b.htm - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://193.6.41.71/ct_release/obj/obj03b.htm
Go Links
Done
Internet
start
Inbox - Outl...
Total Comm...
thesis-0324...
leighton - Mi...
MIE2002cik...
cikk - Micros...
http://193....
EN
11:07

(follow-up scan version)

1hr27min (early scan)

56-year old man

Patient Notes: He had aphasia, and eye deviation.

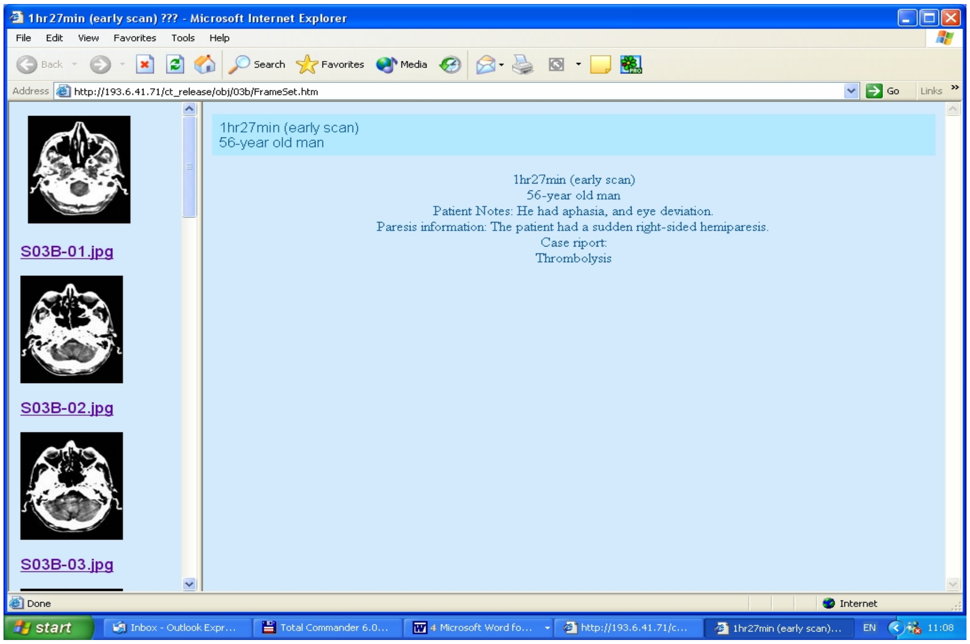
Paresis information: The patient had a sudden right-sided hemiparesis.

Case riport:
Thrombolysis

Pictures of this version, and the link to the other (early) version.

©CIR

Figure 3.9. Result screen of the first hit



1hr27min (early scan)
56-year old man

1hr27min (early scan)
56-year old man
Patient Notes: He had aphasia, and eye deviation.
Paresis information: The patient had a sudden right-sided hemiparesis.
Case riport:
Thrombolysis

Figure 3.10. CT images of the first hit.

3.6.2 Evaluation of NeuRadIR Using the Leighton-Srivastava Method

(Leighton and Srivastava, 1999) elaborated and applied a method to compare five Web search engines (Alta Vista, Excite, HotBot, Infoseek, Lycos) for precision on the first twenty results returned for fifteen queries. Because their method allows for evaluating a real search application on the Web, it has been used to evaluate the precision of NeuRadIR using HIR model as a search method.

Although the application NeuRadIR does not have the complexity of a usual Web search engine (like those above), the method can be adopted based on the formula (Dominich et al, 2001) to compute precision can be developed; these are described in the following. The tests were carried out during March 2004.

3.6.2.1 Search Method

The database was stable in the test period.

The search method used was that of NeuRadIR i.e., the hyperbolic information retrieval with *tfn* weighting schemes. The parameters are the following:

The hyperbolic values are computed using this formula:

$$Hyp = \frac{1}{1 + \ln \dfrac{r + A}{r - A}}, \text{ where } A = \sqrt{\sum_{i=1}^{n} \left(w_{ij} - q_i \right)^2} \qquad (3.20)$$

where radius r is the following (because we suppose, that it is needed a categorical retrieval system):

$$r = \max A + 10^{-10} \tag{3.21}$$

A cutoff – value (cv) was introduced to exclude the hits which does not include any query terms:

$$\text{cv} = \left(\ln\left(e \cdot \frac{r + \max A}{r - \max A} \right) \right)^{-1} \tag{3.22}$$

3.6.2.2 Evaluation Method

A set of criteria (Relevance Categories) was established first, before evaluating any links. The relevance was defined in two different ways supposing two different types of user. Separate searches were performed for each query. The returned hits were evaluated: placement in a relevance category and calculation of numeric precision.

(Xu, 1999) reported that from 1996 to 1999, for more than 70 % of the time, user only views the top ten results. Corresponding to this report the suggested formula of (Dominich at al, 2001) evaluating the first ten hits is used.

3.6.2.3 Relevance Categories

The relevance categories suggested by (Leighton and Srivastava, 1999) are the following:

- Category 0: duplicate links, inactive links (file not found, forbidden, server not responding), irrelevant links
- Category 1: technically relevant links
- Category 2: potentially useful links,
- Category 3: a most probably useful links.

A document is either in a category or not in the category.

The suggested formula for the metric begins by converting the categories into binary values of zero or one. The links in categories 1, 2, and 3 are assigned as one in the formula, and the links in category 0 are assigned zero. The first twenty links are divided into three groups: the first three links (multiplied by 20), the next seven links round out the user's first result page (multiplied by 17), and the last ten (multiplied by 10), make up the second page of result. The formula is the following:

$$\frac{Links_{1\text{-}3} \times 20 + Links_{4\text{-}10} \times 17 + Links_{11\text{-}20} \times 10}{279 - (missing\_links \times 10)} \tag{3.23}$$

In the suggested formula of (Dominich et al, 2001) evaluating the first ten hits the classes and their weights for the first ten precision cases are as follows:

(i) Class 1: contains the first two links and has weight 20,
(ii) Class 2: contains the next three links and has weight 17.
(iii) Class 3: contains the last five links and has weight 10.

Taking the classes into account, and following the line for the first five precision cases the final formula in this case is as follows:

$$\frac{Links_{1,2} \times 20 + Links_{3,4,5} \times 17 + Links_{6\text{-}10} \times 10}{141 - missing\_links \times 10} \qquad (3.24)$$

Because NeuRadIR does not have the complexity of a usual Web search engine, and the documents consists of the database contain only some sentences, fewer categories than those suggested sufficed. They are as follows:

- Category 0: inactive links (file not found, forbidden, server not responding), duplicate links
- Category 1: irrelevant hits
- Category 2: relevant hits, (either technically, i.e., the document contains the search expressions, and/or the document is judged to be relevant due to its content).

The conversion of categories to values was made in two different ways depending on the definition and judgement of relevance:

- *Version A* (traditional, suggested by Leighton and Srivastava) called "rigorous" user-based:
  - Category 0: inactive links (file not found, forbidden, server not responding), duplicate links
  - Category 1: irrelevant hits, meaning all the hits that do not satisfy the complex queries, i.e. they do not contain all the query terms.
  - Category 2: relevant hits meaning all the hits which contain all the query terms.

Thus, the hits of "Category 0", or "Category 1" are assigned zero and the hits of "Category 2" are assigned 1.

- *Version B* (modified Leighton and Srivastava formula) called "permissive" user-based, or partial relevance:

  It is assumed that the users are mostly satisfied with the results containing all of the index terms of the query but they want to see also these results that contain only a part of the query. For that case, a new category was introduced, thus the categories are the following:

- Category 0: inactive links (file not found, forbidden, server not responding), duplicate links
- Category 1: irrelevant hits (all the hits which does not contain any index terms of the query)
- Category 2: partial relevant hits, (all the hits which contains only a part of the query).
- Category 3: relevant hits (all the hits containing all of the index terms of the query, i.e. it contains the complex query)

Thus, the hits of "Category 0", or "Category 1" are assigned zero, and the hits of "Category 2" are 0.5 and the hits of "Category 3" are 1.

3.6.2.4 Test Suite

Because the target users of NeuRadIR are — typically but not necessarily restricted to — medical practitioner doctors, the queries are complex queries, e.g. they want to know the connections between the symptoms and diseases. Additionally, corresponding to the results of (Spink and Xu, 2000) and (Jansen at al, 1998, 2000), that on average, a user query contains 2.21 terms, the test queries of this experiment contained two index terms.

In order to try and minimise biases (it is well known that biases, both conscious and unconscious, do affect any such test to a certain extent, and this cannot be totally excluded), different numbers of verbal requests of — randomly selected — users were recorded, appropriate queries were formulated, and retrievals were performed accordingly in order to establish the exact search expressions and their number.

A number of 30 queries were decided on; fewer queries proved to have a considerable biasing effect, more queries did not yield better results. The precision was computed by the suggested formula of (Dominich et al, 2001) in the form of (3.23).

3.6.2.5 Results and Discussion

The results of the experiments are summarized in table 3.6 using version "A", and "B" to compute the precision. Every row assigns a query. The columns give the number of the retrieved hits, the relevant, irrelevant and partial relevant hits and the precision of the method for a given query, and for the version "A", and "B". The values of the average precision can be seen in the last row. It can be clearly seen that the NeuRadIR application based on HIR model — with the precision of 0.578, and 0.77 — meets very well users' satisfaction.

Table 3.6. Results of the experiments version A, and B.

		Version A			**Version B**			
Query	Number of hits	Relevant hits	Irrelevant hits	Precision	Relevant hits	Partial relevant hits	Irrelevant hits	Precision
1.	16	1-4	5-16	0.525	1-4	5-16	-	0.762
2.	18	-	1-18	0	-	1-18	-	0.5
3.	27	1-6	7-27	0.716	1-6	7-27	-	0.858
4.	16	1-2	3-16	0.284	1-2	3-16	-	0.642
5.	18	1-2	3-18	0.284	1-2	3-18	-	0.642
6.	25	1-10	11-20	1	1-10	11-20	-	1
7.	33	1-11	12-30	1	1-11	12-30	-	1
8.	20	1-10	11-20	1	1-10	11-20	-	1
9.	20	0	1-20	0	0	1-20	-	0.5
10.	28	1-12	13-28	1	1-12	13-28	-	1
11.	16	1-2	3-16	0.284	1-2	3-16	-	0.642
12.	14	1-2	3-14	0.284	1-2	3-14	-	0.642
13.	26	1-5	6-26	0.645	1-5	6-26	-	0.823
14.	16	0	1-16	0	0	1-16	-	0.5
15.	16	1-2	3-16	0.284	1-2	3-16	-	0.642
16.	26	1-7	8-26	0.787	1-7	8-26	-	0.894
17.	34	1-8	9-34	0.858	1-8	9-34	-	0.929
18.	18	0	1-18	0	0	1-18	-	0.5
19.	28	1-10	11-28	1	1-10	11-28	-	1
20.	26	0	1-26	0	0	1-26	-	0.5
21.	28	0	1-28	0	0	1-28	-	0.5
22.	31	1-12	13-31	1	1-12	13-31	-	1
23.	26	1-2	3-26	0.284	1-2	3-26	-	0.642
24.	28	1-2	3-28	0.284	1-2	3-28	-	0.642
25.	33	1-12	13-33	1	1-12	13-33	-	1
26.	34	1-20	21-34	1	1-20	21-34	-	1
27.	26	1-4	5-26	0.525	1-4	5-26	-	0.762
28.	36	1-14	15-36	1	1-14	15-36	-	1
29.	35	1-17	18-35	1	1-17	18-35	-	1
30.	29	1-3	4-29	0.4	1-3	4-29	-	0.702
Average				**0.548**				**0.77**

CHAPTER 4

VARYING RETRIEVAL CATEGORICITY

In this chapter, the concept of entropy is used to define retrieval categoricity. Then, retrieval categoricity of the VSM and the HIR model is investigated, and a new efficient way to vary retrieval categoricity is introduced.

4.1 Motivation

Claude Shannon, in his classical paper (Shannon, 1948), defined the concept of information as one's freedom of choice (to select from alternatives). He also introduced a measure for information which has maximum value when one has total freedom of choice, and has minimal value when has no freedom in selection. In other words, when it is known exactly what to select then uncertainty is decreased, but when we are free to choose any alternative we want then uncertainty increases. The concept of and formula for entropy has been used in information retrieval in a number of ways.

As early as 1969 (Meetham, 1969), and somewhat later in (Guazzo, 1977), the concepts of entropy and Shannon information have been applied to IR evaluation as better alternatives to precision and recall. In the 1980's, the maximum entropy principle (MEP) was applied to IR (Cooper and Huizinga, 1982; Kantor, 1984). Formally, MEP can be expressed as a constrained optimisation problem, in which one wishes to determine the probability distribution associated to a random variable over a discrete space which has the greatest entropy subject to constraints (these express the knowledge that we impose upon this distribution).

In IR, MEP can be formulated as follows:

Let an elementary event w denote the observation of a document with respect to a given query. The probability $p(w)$ of an event depends on whether the document is relevant or not, and on whether it contains query terms or not.

The retrieval system aims at maximizing the associated entropy:

$$\sum_{w} p(w) \cdot \log p(w)$$

is subject to constraints (such as, e.g., the probabilities of relevant/non-relevant documents to contain query terms).

MEP proves useful as a formal research tool; (Greiff and Ponte, 1998) show that MEP can be applied as a formal framework in which the probabilistic IR model (Robertson and Sparck Jones, 1977) can be obtained. At the same time, it seems to be less effective when applied to retrieval in practice: extensive experiments show that MEP works well for small document collections but seems to be progressively worse for larger ones (Kantor and Lee, 1998). However, MEP proved useful in text classification tasks as shown by experiments carried out in (Nigam, Lafferty and McCallum, 1999). Entropy has been applied to other IR tasks as well. In (Fujii and Ishikawa, 2001) the associated entropy — where C denotes a cluster of documents, and $p(C)$ the probability that a relevant document belongs to cluster C — can be used as a measure of the clustering process. Let $p(y)$ denote the probability of a word y as its frequency (i.e., its count over total number of words). Then, a measure of the reduction in uncertainty about whether the word y will be the next word in a sequence of text (given that x was the previous word) can be expressed by entropy (Berger and Lafferty, 1999). In (Yoo et al., 2002) etropy is used for texture modelling in image indexing and retrieval. The information content of a collection of documents consisting of — not necessarily disjoint — classes is the entropy associated to class cardinalities, which is being reduced, in the retrieval process, from its maximum (Baclawski and Simovici, 1996). (Tan et al., 2002) have used MEP for text categorisation, and showed that the use of bigrams in addition to single words can increase performance.

It can hence be seen that entropy (Shannon information) has been used to formalise the probabilistic IR model, to construct practical retrieval systems, to cluster documents, to model texture in image retrieval. My dissertation aims at applying it for a different purpose. The concept of entropy is used to define an amount of uncertainty U associated with answers in the Vector Space Model of information retrieval, and to define the connected concept categoricity. Based on this concept the retrieval categoricity of the VSM — equipped different similarity measures and weighting schemes — and of the HIR model is investigated and a new effective way to vary retrieval categoricity is introduced in this chapter. In *Chapter 5* it will be shown that a retrieval system using positive RSV (Retrieval Status Value) for retrieval may be conceived as a probability space in which the quantity of the associated amount of Shannon information is being reduced. This result will be applied to the calculation of term discrimination values.

4.2 Information Theory and Entropy

In this section, in order to fix the ideas, the concept of a probability space is briefly recalled (Kolmogoroff, 1933). This will be followed by a short review of the concept of Shannon (1948) information and its measure.

4.2.1 Probability Space

In order to fix the ideas in *Chapter 4* and *5*, the concept of a probability space is briefly recalled (Kolmogoroff, 1933):

Given a set Ω called *universe*; let its elements be called elementary *events*.

A set $\Im \subseteq \wp(\Omega)$ is called a σ-algebra if $\Omega \in \Im$, and

$$A \cap B \in \Im, \quad A \cup B \in \Im, \quad \Omega \setminus A \in \Im, \ \forall A, B \in \Im$$

A *probability measure* is a function P: $\Im \to [0; 1]$ satisfying the following properties:

$$P(\Omega) = 1; \quad A \cap B = \varnothing \Rightarrow P(A \cup B) = P(A) + P(B); \quad \forall A, B \in \Im$$

The triple $(\Omega, \Im, P) = \Psi$ is called a *probability space*.

4.2.2 Amount of Shannon Information

Shannon's creation (1948) of the subject of information theory is one of the great intellectual achievements of the twentieth century. Information theory has had an important and significant influence on mathematics, particularly on probability theory, but he did his work primarily in the context of communication engineering. The formula proposed as a measure for information is an expression of the quantity of Shannon information, which is called entropy. While entropy is interpreted as a measure of uncertainty, information is viewed as a reduction in the level of uncertainty. Thus, the amount of Shannon information may also be viewed as a certain expression of uncertainty level, they can, in principle, be used as equivalent concepts; when it is known exactly what to select then uncertainty is decreased, but when we are free to choose any alternative we want then uncertainty is highest.

The definition (Shannon, 1948) of a measure for information is the following:

Given events (alternatives) E_j, $j = 1,\dots, m \in$ **N** (**N** denotes the set of natural numbers);

let p_j denote the probability to select alternative (probability of occurrence of the event) E_j.

A measure H for information is defined as follows:

$$H = -k \cdot \sum_{j=1}^{m} p_j \cdot \log_2 p_j \,,$$

where k is a positive constant, which amounts to a choice of a unit of measure; in what follows k will be taken as being equal to 1. Thus,

$$H = -\sum_{j=1}^{m} p_j \cdot \log_2 p_j \,, \tag{4.1}$$

The quantity H satisfies the following properties:

- The amount of information is zero if and only if exactly one alternative is selected:

$$\lim_{\substack{p_k \to 1 \\ p_j \to 0, \forall j \neq k}} H = 0 \qquad \Leftrightarrow \qquad (\, p_k = 1; \quad p_j = 0, \quad \forall j \neq k \,) \tag{4.2}$$

- The amount of information is maximal and equal to $\log_2 m$ if all p_j are equal to $1/m$:

$$(\, p_j = \frac{1}{m}, \quad j = 1, \ldots, m \,) \qquad \Rightarrow \qquad H = \max H = \log_2 m \tag{4.3}$$

- The farther apart p_j from each other the smaller the amount of information:

$$P < P' \qquad \Rightarrow \qquad H > H' \tag{4.4}$$

where the deviation of the probabilities p_j from each other can be defined in the following form of P:

$$P = \sum_{j=1}^{m} \left(\frac{1}{m} - p_j \right)^2 \text{, and } P' = \sum_{j=1}^{m} \left(\frac{1}{m} - p'_j \right)^2 ,$$

The property (4.4) will play an important role in the present paper, thus, without restricting its general validity, let us have a closer look at the case $m = 2$.

THEOREM 4.1. $P < P' \Rightarrow H > H'$.

***Proof*:**

Let $p_1 = p$, $p_2 = q$. The condition $P < P'$ means that

$$(0.5 - p)^2 + (0.5 - q)^2 < (0.5 - p')^2 + (0.5 - q')^2.$$

We can assume that $p' = p + a$, $q' = q - a$. From this we obtain $p - q + a > 0$. From

$$q - a < p \quad \text{and} \quad \log(1/(q - a)) > \log(1/p)$$

it follows that

$$p \cdot \log(1/p) > (q - a) \cdot \log(1/(q - a));$$

whereas from

$$q < p + a \quad \text{and} \quad \log(1/q) > \log(1/(p + a))$$

it follows that

$$q \cdot \log(1/q) > (p + a) \cdot \log(1/(p + a)).$$

Hence $p \cdot \log(1/p) + q \cdot \log(1/q) > p' \cdot \log(1/p') + q' \cdot \log(1/q')$, i.e., $H > H'$.♦

In other words, if the probabilities p_j change such that they deviate more from $1/m$ (some are closer to 1 while the others closer to zero) the amount of information (or uncertainty) becomes smaller, i.e., the freedom to select becomes more restricted.

Thus, for example, if one has total freedom to choose from two alternatives then the amount of information associated to this situation is considered to be unity (i.e., 1 bit).

4.3 Uncertainty Decreasing Operation (UDO)

The meaning of the property (4.4) in other words is the following: if the probabilities p_j change such that they deviate more from $1/m$ (some are closer to 1 while others are closer to zero) the amount of information (or uncertainty) becomes smaller, i.e., the freedom to select becomes more restricted. This entitles us to introduce the following:

DEFINITION 4.1 An operation (procedure, process, mechanism), which spreads the probabilities p_j, is called an *Uncertainty Decreasing Operation* (UDO).♦

Thus, any operation that constrains the freedom (and thus reduces the uncertainty) to select is an UDO.

EXAMPLE 4.1

(i) Numerical minimisation of a function based on the gradient method: the freedom to select a direction to follow is decreased because only that given by the gradient can be followed.

(ii) Breadth-first search algorithm: the freedom to move to a next vertex is constrained because a downward walk is not allowed as long as there are unexplored breadth vertices.

(iii) A person on diet does not (or should not) have total liberty to choose the bread or meat he/she would like, so far his/her uncertainty in choosing foods are decreased.

In cases like these, the freedom of choice to select from alternatives is constrained, some of the alternatives are/should be selected with higher probabilities in the detriment of the others, and thus the total amount of information (and uncertainty) associated to the selection situation as a whole is decreased.

4.4 Retrieval Status Value-based Retrieval as Uncertainty Decreasing Operation [THESIS 2.a]

Information retrieval systems typically rank documents according to their "retrieval status values" (RSV) with respect to a given query at a given time. The higher the RSV of a document for a given query means the higher chance of the document to be selected as an answer.

EXAMPLE 4.2

(i) In the Boolean Model RSVs are either zero ore one. A document with a RSV-value of "one" means that the document is a hit for a given query; in the other hand, a document with "zero" RSV-value means that the document is not a hit for a given query.

(ii) Fuzzy retrieval allows for RSVs in the interval [0,1]. For a given query a document with 0.75 RSV precedes the document with 0.25 RSV-value in the ranked hit list.

Based on the concepts of a probability space and UDO, it is shown in [P2] that any retrieval model or system based on positive RSV — e.g., vector space, probabilistic, Boolean, coordination level matching, fuzzy, connectionist interaction, link analysis retrieval models (Dominich, 2001) — may be conceived as a probability space that decreases the amount of the associated Shannon information, i.e., it is an UDO probability space.

It is first proved that a probability space having its probability measure defined in a certain way (normalised so that its values sum up to unity) is an UDO.

LEMMA 4.1 Let Ψ = (Ω, $\Im$, P), $|\Omega|$ = m, denote a probability space with the probability measure P defined as follows:

$$P(X) = \begin{cases} \dfrac{r_j}{\sum_{k=1}^{m} r_k}, X = X_j \in \Omega \\ \\ P'(X), otherwise \end{cases} \tag{4.5}$$

where not all r_j are equal to each other, i.e., $\exists$ $k \neq s$ such that $r_k \neq r_s$. (An explicit formula for *P'* does not play any role in this context. It was defined only pro forma) Then the probability space Ψ is an UDO.

***Proof*:**

By assumption, not all r_j are equal to each other,

i.e., $\exists$ $k \neq s$ such that $r_k \neq r_s$ in other words:

$$\sum_{\substack{k=1\\ s=k+1}}^{\substack{m\\ m-1}} \left(r_k - r_s\right)^2 > 0,$$

which is equivalent to the following:

$$\sum_{\substack{k=1\\ s=k+1}}^{\substack{m\\ m-1}} \left(r_k - r_s\right)^2 = \left(m-1\right)\cdot \sum_{k=1}^{m} r_k{}^2 - 2\cdot \sum_{\substack{k=1\\ s=k+1}}^{\substack{m\\ m-1}} \left(r_k - r_s\right)^2 > 0,$$

in other form: $\left(m-1\right)\cdot \sum_{k=1}^{m} r_k{}^2 > 2\cdot \sum_{\substack{k=1\\ s=k+1}}^{\substack{m\\ m-1}} \left(r_k - r_s\right)^2$

The space Ψ is an UDO if it spreads the probabilities from 1/*m* (Def. 4.1), i.e., we have to show that

$$P = \sum_{j=1}^{m} \left(\frac{1}{m} - p_j\right)^2 > 0, \qquad \text{where } p_j = \frac{r_j}{\sum_{k=1}^{m} r_k}.$$

We can write *P* in the following form:

$$P = \sum_{j=1}^{m} \left(\frac{1}{m} - p_j\right)^2 = \sum_{j=1}^{m} \left(\frac{1}{m} - \frac{r_j}{\sum_{k=1}^{m} r_k}\right)^2 = \sum_{j=1}^{m} \left(\frac{1}{m^2} - \frac{2}{m}\frac{r_j}{\sum_{k=1}^{m} r_k} + \frac{r_j{}^2}{\left(\sum_{k=1}^{m} r_k\right)^2}\right) =$$

$$= \sum_{j=1}^{m} \frac{1}{m^2} - \sum_{j=1}^{m} \frac{2}{m}\cdot\frac{r_j}{\sum_{k=1}^{m} r_k} + \sum_{j=1}^{m} \frac{r_j{}^2}{\left(\sum_{k=1}^{m} r_k\right)^2} = \frac{1}{m} - \frac{2}{m} + \sum_{j=1}^{m} \frac{r_j{}^2}{\left(\sum_{k=1}^{m} r_k\right)^2} =$$

$$= \sum_{j=1}^{m} \frac{r_j{}^2}{\left(\sum_{k=1}^{m} r_k\right)^2} - \frac{1}{m} = \frac{\sum_{j=1}^{m} r_j{}^2}{\left(\sum_{k=1}^{m} r_k\right)^2} - \frac{1}{m}$$

Because, $\left(m-1\right)\cdot \sum_{k=1}^{m} r_k{}^2 > 2\cdot \sum_{\substack{k=1\\ s=k+1}}^{\substack{m\\ m-1}} \left(r_k - r_s\right)^2$,

$$\left(\sum_{k=1}^{m} r_k\right)^2 = \sum_{k=1}^{m} r_k{}^2 + 2\cdot \sum_{\substack{k=1\\ s=k+1}}^{\substack{m\\ m-1}} \left(r_k - r_s\right)^2 < \sum_{k=1}^{m} r_k{}^2 + \left(m-1\right)\cdot \sum_{k=1}^{m} r_k{}^2 = m\cdot \sum_{k=1}^{m} r_k{}^2 \;,$$

it follows, that

$$1 < \frac{m\cdot \sum_{k=1}^{m} r_k{}^2}{\left(\sum_{k=1}^{m} r_k\right)^2} \text{, which means, that } 0 < \frac{\sum_{j=1}^{m} r_j{}^2}{\left(\sum_{k=1}^{m} r_k\right)^2} - \frac{1}{m} \qquad \blacklozenge$$

It is proved that the probability space defined in (4.5.) is an UDO (the freedom of choice to select from alternatives is constrained), because it spreads the probabilities.

Now it can be shown that any RSV-based retrieval system can be conceived as a probability space that decreases the amount of information.

THEOREM 4.2. Any positive RSV-based retrieval system is an UDO probability space Ψ.

***Proof*:**

Given an RSV-based retrieval system.

Let r_j denote the RSV of document D_j relative to query Q. In other words, r_j may be viewed as representing a degree of the choice of document D_j as a response to query Q. The higher the value of r_j the higher the chance of document D_j to be selected as an answer.

A sequence $\langle \mathrm{r} \rangle = r_j, \ldots, r_m$ can be defined, which represents the choices of all documents relative to query Q. (The no-hit case, i.e., when all the r_j are null, can be excluded as trivial.) Using the sequence $\langle \mathrm{r} \rangle$, a sequence $\langle \boldsymbol{P} \rangle = p_1, \ldots, p_m$, where

$$p_j = \frac{r_j}{\sum_{k=1}^{m} r_k}, \; j = 1,\ldots, m$$

is defined, which can be viewed as the probabilities to select the documents as answers in the following probability space

$$\Psi = (\Omega, \Im, P), \qquad \Omega = \{D_1, \ldots, D_m\},$$

$$P(D_j) = p_j, \qquad P(X) = P'(X) \text{ if } X \neq D_j.$$

Hence (Lemma 4.1), the positive RSV-based retrieval system —because it can be viewed as a probability space — is an UDO.♦

4.5 Categoricity

Using the concept of Shannon amount of information (4.1), an amount of uncertainty U associated with answers in VSM was defined as:

$$U = -\sum_{j=1}^{m} p_j \log_2 p_j \ ,$$

where probabilities p_j are in the following form:

$$p_j = \frac{r_j}{\sum_{k=1}^{m} r_k} \ j = 1, \ldots, m \tag{4.6}$$

r_j denotes the RSV (retrieval status value) of document D_j relative to query Q.

EXAMPLE 4.3. The following example illustrates the computation of uncertainty U. Let us consider three documents: D_1, D_2 and D_3, $m = 3$, two terms: t_1 and t_2, $n = 2$.

Let the *term-by-document* matrix be as follows: $W = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 3 & 2 \end{bmatrix}$.

Let Q denote a query and the corresponding term frequencies be (0, 1).

For computational convenience, matrix notation is used.

- If the retrieval function r is the dot product, then the chances r_1, r_2 and r_3 are

  $W^{\mathrm{T}}\mathbf{q} = \begin{bmatrix} 2 & 0 \\ 1 & 3 \\ 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \\ 2 \end{bmatrix}$, while the corresponding probabilities are $\begin{bmatrix} 0 \\ 0.6 \\ 0.4 \end{bmatrix}$.

  The associated amount of information is decreased to 0.971 from the maximum $\log_2 3 = 1.585$.

- If the retrieval function is the Cosine measure and the weights are max normalised, the chances r_1, r_2 and r_3 are $\begin{bmatrix} 0 \\ 0.923 \\ 0.847 \end{bmatrix}$,

  whilst the corresponding probabilities are $\begin{bmatrix} 0 \\ 0.521 \\ 0.479 \end{bmatrix}$.

The associated amount of information is decreased to 0.999 from the maximum $\log_2 3 = 1.585$.

A concept of categoricity has been introduced connected to the uncertainty.

DEFINITION 4.2. Let "A" and "B" denote two different and positive RSV-based retrieval methods; U_A, and U_B denote the uncertainty associated with answers. Retrieval system based on the "A" method is *more categorical*, than "B" in its answers, if $U_A < U_B$.

Thus, a retrieval system is categorical in its answers if it decreases uncertainty, so, it is known exactly what to select from the answers. In other words categoricity means the spreading of the answers' relevance values.

EXAMPLE 4.4. The following example illustrates uncertainty, categoricity and the parallel between them. Using the Google Search with its Page Rank toolbar, the experiment with two different queries was carried out on the 2nd of January 2004. "Seat" and "North Sea" query terms were issued. D_j denotes the j-th document in the ranked hit list, and r_j its relevance values for a given query. The results are the following:

Case 1.: answers (D_j) and their relevance values (r_j) for the query "Seat", as follows:

$(D_2;\ r_2 = 0.5)$, $(D_5;\ r_5 = 0.5)$, $(D_9;\ r_9 = 0.5)$.

Case 2.: answers (D_j) and their relevance values (r_j) for the query "North Sea", as follows:

$(D_1;\ r_1 = 0.7)$, $(D_2;\ r_2 = 0.6)$, $(D_3;\ r_3 = 0.5)$.

In the first case the answers are less categorical, they have the same relevance values, so the user is more uncertain than in the second case.

4.6 Method for the Study of the Relationship between Entropy Reduction, Weighting Scheme and Similarity Measure [THESIS 2.b]

Based on Theorem 4.2 — any positive RSV-based retrieval system is an UDO probability space — a method is developed in [P3] for the practical study of the relationship between entropy reduction, weighting scheme, and similarity measure in RSV-based retrieval systems. The first step of the method is the implementation of the IR model, and after the query formulation the entropy and the entropy reduction can be computed. The second step is the study of the RSV-based IR model using different weighting schemes, and similarity measures.

4.6.1 Steps of the method

Method for the study of the relationship between entropy reduction, weighting scheme, and similarity measure is the following:

Step 1. Computation of categoricity as entropy reduction:

Given the following:

- a database with a fixed number of index terms, documents,
- a query,
- a weighting scheme,
- an RSV computation method.

1.1. Compute the maximum value of entropy as U_{max} = log *m* (formula 4.3).

1.2. Implement the IR model under focus:

1.2.1. Generate the *term-by-document* matrix.

1.2.2. Formulate a query *Q*.

1.2.3. Compute the RSV-values.

1.3. Compute entropy *U* (formula 4.6).

1.4. Calculate entropy reduction as $U_{max} - U$ (alternatively as %).

Step 2. Study of RSV-based IR model:

Given the following:

- a database with a fixed number of index terms, documents,
- a query,
- different weighting scheme,
- different RSV computation methods.

2.1. Perform step 1.2 –1.3. for each weighting scheme and RSV computation method.

2.2. Calculate average values of entropy *U* for each weighting scheme and RSV computation method.

2.3. Calculate entropy reduction as U_{max} – average values of *U* (alternatively as %) for each weighting scheme, and RSV computation method.

2.4. Create a table with the results obtained.

2.5. Draw conclusions.

4.6.2 Experiments

Based on the method described in section 4.6.1 two experiments — using the medical database of the NeuRadIR (section 2.2.4) — are performed to study of the entropy reduction, and thus the categoricity property of the VSM model with different similarity measures (Cosine, Dice, Jaccard, Dot) and weighting schemes (*frequency, maxNorm, term-frequency normalized*):

Case 1.: Binary local term weights ($\chi(f_{ij})$) were used as weighting scheme,

Case 2.: Term occurrences (f_{ij}) were used as weighting scheme.

Whilst the medical database contains 68 index terms, 40 documents, and six different weighting schemes were used, hence six 68×40 ***term-by document*** matrices were generated using Mathcad, thus the step 1.2. of the method was repeated six times. Step 1.3 was repeated 24 times (because of six weighting scheme, and four RSV computation method). Step 2.2, and step 2.3 were repeated for all weighting scheme, and RSV computation method, thus 14 times. The computation was performed using Mathcad. The table — step 2.4 —, which shows the results of the experiments, can be found in table 4.1. It shows the amount of information ***U*** in case of four retrieval measures with three weighting schemes — using two local terms weights — and the percentages of the decrement of ***U*** in every case over the medical database.

Table 4.1. Decrement of the amount of information in VSM over the Medical Database of the NeuRadIR.

Weighting scheme	**Amount of information *U* if retrieval measure is:**				**Average values of *U***	***U*** $_{max}$	***U* decreased by (%)**
	Cosine	***Dice***	***Jaccard***	***Dot***			
Medical Database using binary local term weights							
frequency $\chi(f_{ij})$	5.151	5.15	5.043	5.128	5.118	5.322	3.8
maxNorm $\frac{c(f_{ij})}{\max_k c(f_{kj})}$	5.151	5.15	5.043	5.128	5.118		3.8
tfn $\frac{c(f_{ij})}{\sqrt{\sum_{i=1}^{n} c(f_{ij})^2}}$	5.151	5.159	5.15	5.151	5.153		3.2
Average values of *U*	5.1151	5.153	5.079	5.136			
***U* decreased by (%)**	3.3	3.3	4.6	3.5			
Medical Database using term frequency (occurrences) local term weights							
frequency f_{ij}	5.161	5.16	5.058	5.134	5.128	5.322	3.7
maxNorm $\frac{f_{ij}}{\max_k f_{kj}}$	5.161	5.145	5.03	5.097	5.108		4
tfn $\frac{f_{ij}}{\sqrt{\sum_{i=1}^{n} f_{ij}^2}}$	5.161	5.169	5.161	5.161	5.163		3
Average values of *U*	5.161	5.158	5.083	5.13			
***U* decreased by (%)**	3	3.1	4.5	3.6			

So, it is shown experimentally that the quantity *U* varies depending on the similarity measure (Cosine, Dot product, Jaccard's and Dice's coefficient) and weighting scheme (*frequency, maxNorm* and *tfn*) used. The experiments show the following results:

(i) The usage of the Cosine/Dice similarity measures resulted in the least entropy reduction for every weighting scheme.

(ii) The usage of the *tfn* (normalised frequency) weighting scheme reduces entropy to the greatest extent for every similarity measure.

(iii) The usage of the Jaccard similarity measure reduces entropy to the greatest extent for every weighting scheme.

(iv) The usage of the *maxNorm* scheme yielded the least entropy reduction for every similarity measure and collection of text.

From the point of view of the user, a parallel can be drawn between uncertainty and categoricity. Uncertainty appears as a degree to which the retrieval system is categorical in its answers: it is less categorical for the Cosine/Dice measure and *tfn* scheme, and most categorical for Jaccard's coefficient with *maxNorm* scheme. In other words, such a VSM is less categorical as regards its answers.

Albeit in the VSM ranking determines the sequence order of answers, users may prefer categorical answers as well, as these may be more convincing and they help the user select answers. However, in the VSM the only way to modify it is to take a different weighting scheme and/or similarity measure. Unfortunately, the rank order and the same answer set cannot be guaranteed.

Section 4.7, and 4.8 investigates that in the hyperbolic similarity measure-based (it was introduced in chapter 3) information retrieval system the categoricity can be varied with less computation, than in the VSM by just varying the radius of the space.

4.7 Radius of Space as a Control Variable for Categoricity [THESIS 2.c]

It is investigated in [P1] that in the hyperbolic similarity measure-based information retrieval system the categoricity can be varied by just varying the radius of the space.

4.7.1 Radius of the Hyperbolic Space

It is shown that in Hyperbolic Information Retrieval Model the higher the radius ***r*** of the C-KHS the closer the hyperbolic similarity ***Hyp*** to unity, i.e.

$$\lim_{r\to\infty} Hyp(w,q) = \left(\ln\left(e \cdot \frac{r + \sqrt{\sum_{i=1}^{n}\left(w_{ij} - q_i\right)^2}}{r - \sqrt{\sum_{i=1}^{n}\left(w_{ij} - q_i\right)^2}} \right) \right)^{-1} = (\ln(e))^{-1} = \ln e = 1$$

This means that the higher the radius ***r*** the closer the amount of uncertainty ***U*** to its maximum, because the "distance" between the answers are decreased (the relevance degrees of the answers are very close to each other). In other words, increasing the radius of the hyperbolic space yields a less categorical retrieval system and conversely: decreasing the radius leads to more categorical answers.

How much can the radius be reduced?

It it shown first that the categoricity of HIR can be varied by changing the radius *r* between the following limits:

$$r = \max_D d_E(QD) + \varepsilon,$$

where $0 < \varepsilon << +\infty$, and

"<<" symbol means "much less than"

In theory, if *r* becomes equal to $\max_D d_E(QD)$, i.e., the document-vector finds itself on the boundary of the space, the document simply 'disappears' in the infinity of the hyperbolic space as shown by the following derivation based on (3.14):

$$\mathrm{D}\rightarrow\mathrm{V} \Rightarrow \lim_{D\rightarrow V} d_H(QD) = +\infty \Rightarrow$$

$$\lim_{D\rightarrow V} Hyp(QD) = \lim_{D\rightarrow V} (\ln(e \cdot \frac{r + d_E(QD)}{r - d_E(QD)}))^{-1} = \lim_{D\rightarrow V} 1 - \frac{d_H(QD)}{1 + d_H(QD)} = 0$$

Hence, changing its radius *r* between the following limits:

$$\max_D d_E(QD) + \varepsilon < r < +\infty$$

the categoricity of the answers can be varied.

So increasing the radius of the hyperbolic space yields less categorical retrieval system and the answers are more categorical if the radius becomes close to the $d_E(QD)$.

4.7.2 Experimental Results

Experiments were carried out on a part of the medical database (section 2.2.4) to demonstrate, that in HIR, changing the radius of the space can vary categoricity of the answers.

Ten documents (from document 7b to document 12a) and ten index terms (from t_{28} to t_{37}) — with the connected weights (from w_{28} to w_{37}) — were used for the experiments. The ***term-by-document matrix*** (table 4.2), and the ***term-by-query vector*** (figure 4.1) using ***tfn*** (term frequency normalised) weighting scheme were computed by Mathcad 2001i Professional Software.

Table 4.2. *Term-by-document* matrix using *tfn* weighting scheme.

	D$_{7b}$	**D$_{8a}$**	**D$_{8b}$**	**D$_{9a}$**	**D$_{9b}$**	**D$_{10a}$**	**D$_{10b}$**	**D$_{11a}$**	**D$_{11b}$**	**D$_{12a}$**
w$_{28}$	0.408	0.000	0.000	0.000	0.500	0.000	0.000	0.000	0.577	0.000
w$_{29}$	0.408	0.500	0.707	0.447	0.500	0.447	0.447	0.577	0.577	0.577
w$_{30}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
w$_{31}$	0.408	0.500	0.000	0.447	0.500	0.447	0.447	0.000	0.000	0.577
w$_{32}$	0.000	0.000	0.000	0.000	0.000	0.447	0.000	0.000	0.000	0.000
w$_{33}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
w$_{34}$	0.408	0.500	0.000	0.447	0.000	0.447	0.447	0.577	0.000	0.577
w$_{35}$	0.408	0.000	0.000	0.000	0.000	0.000	0.447	0.000	0.000	0.000
w$_{36}$	0.000	0.000	0.000	0.447	0.000	0.000	0.000	0.000	0.000	0.000
w$_{37}$	0.408	0.500	0.707	0.447	0.500	0.447	0.447	0.577	0.577	0.000

$$q = \begin{pmatrix} 0.447 \\ 0 \\ 0.447 \\ 0.447 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.447 \\ 0.447 \end{pmatrix}$$

Figure 4.1. *Term-by-query* vector using *tfn* weighting scheme.

The computation of the uncertainty *U* and the similarity values — cosine, and hyperbolic — were performed on Mathcad.

The uncertainty *U* was computed in the following form:

$$U = -\sum_{j=1}^{m} p_j \log_2 p_j = -\sum_{j=1}^{m} \frac{s_j}{\sum_{j=1}^{m} s_j} \log_2 \left(\frac{s_j}{\sum_{j=1}^{m} s_j} \right),$$

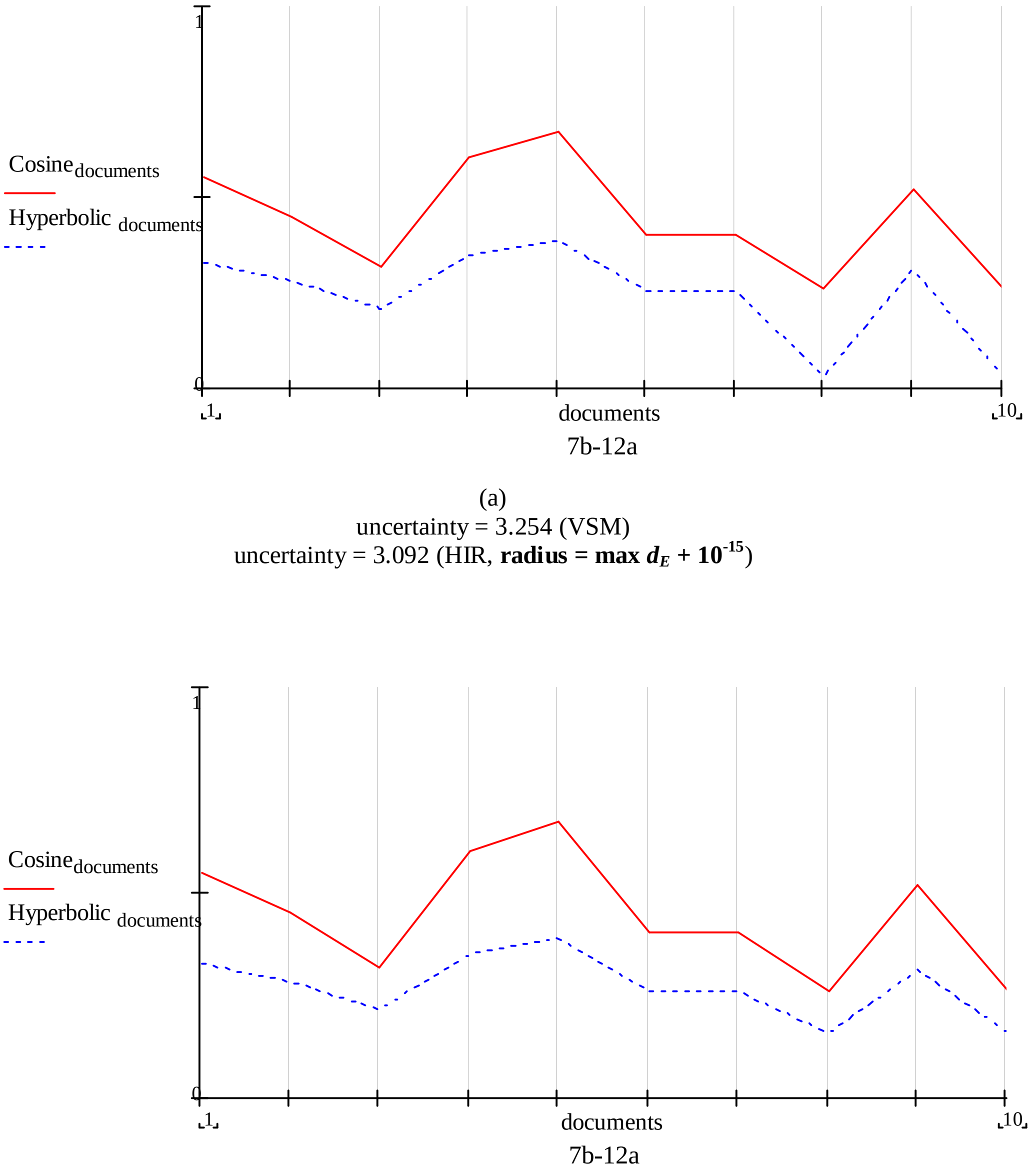
where s_j denotes the similarity — Hyperbolic, and Cosine — measure.

Table 4.3 shows the similarity values of Hyperbolic measure with 4 different radii and Cosine measure.

Table 4.3. Hyperbolic similarity values with different radii, and uncertainty of Cosine and Hyperbolic measures.

Cosine measure uncertainty = 3.254 (VSM)										
	D$_{7b}$	**D$_{8a}$**	**D$_{8b}$**	**D$_{9a}$**	**D$_{9b}$**	**D$_{10a}$**	**D$_{10b}$**	**D$_{11a}$**	**D$_{11b}$**	**D$_{12a}$**
	0.548	0.447	0.316	0.600	0.671	0.400	0.400	0.258	0.516	0.258
Hyperbolic measure										
	D$_{7b}$	**D$_{8a}$**	**D$_{8b}$**	**D$_{9a}$**	**D$_{9b}$**	**D$_{10a}$**	**D$_{10b}$**	**D$_{11a}$**	**D$_{11b}$**	**D$_{12a}$**
r= max d_E + 10^{-15} uncertainty = 3.092	0.323	0.277	0.204	0.348	0.383	0.254	0.254	0.029	0.309	0.029
r= max d_E + 10^{-2} uncertainty = 3.264	0.326	0.281	0.212	0.351	0.386	0.259	0.259	0.154	0.312	0.154
r= max d_E + 1 uncertainty = 3.318	0.522	0.492	0.460	0.539	0.566	0.480	0.480	0.448	0.512	0.448
r= max d_E + 100 uncertainty = 3.322	0.982	0.980	0.977	0.983	0.984	0.979	0.979	0.976	0.981	0.976

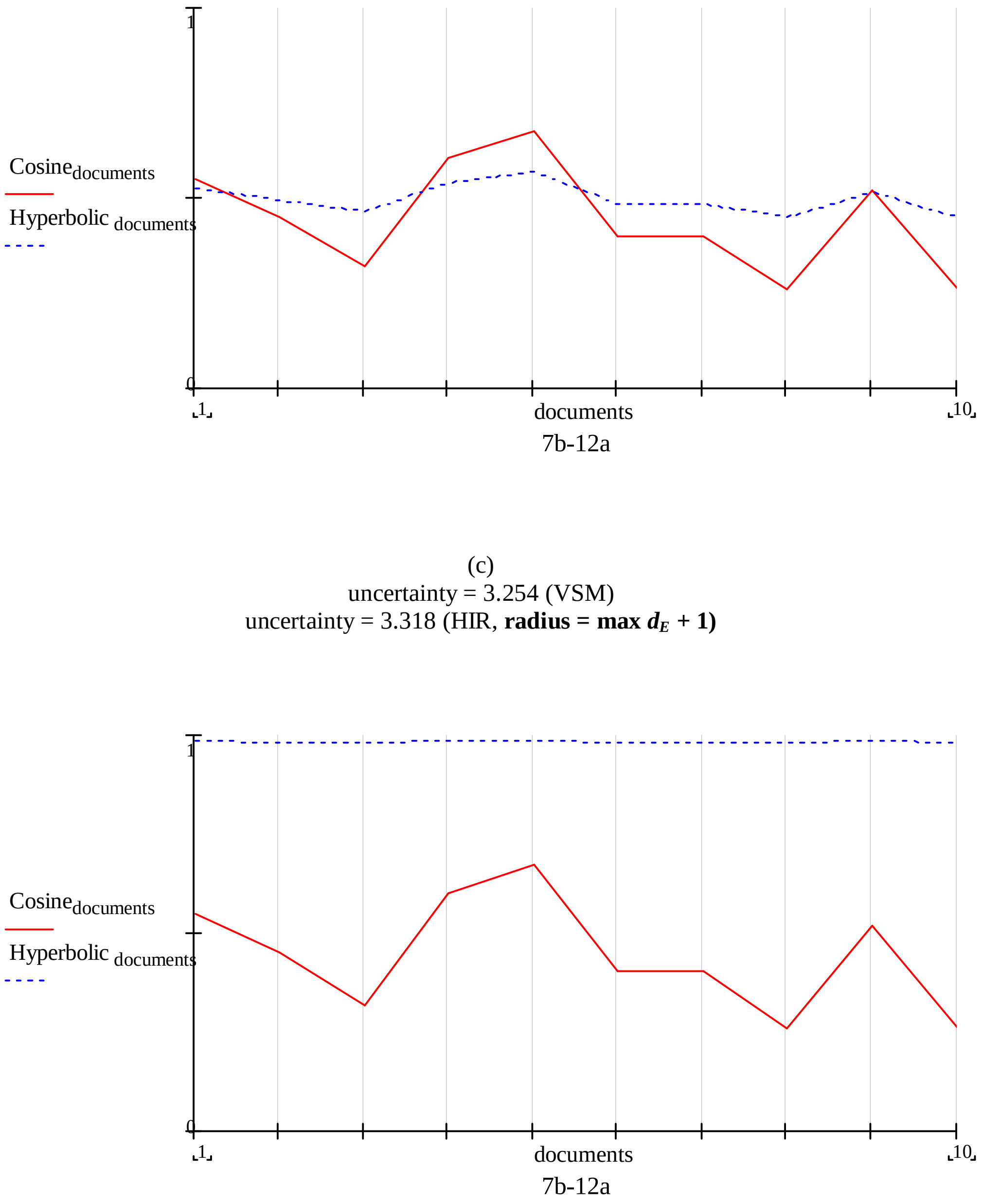
Table 4.3 and figure 4.2 illustrates the modification of categoricity as a function of radius in two different forms.



(a)
uncertainty = 3.254 (VSM)
uncertainty = 3.092 (HIR, **radius = max** $\boldsymbol{d_E}$ **+ 10**$^{-15}$)

(b)
uncertainty = 3.254 (VSM)
uncertainty = 3.264 (HIR, **radius = max** $\boldsymbol{d_E}$ **+ 10**$^{-2}$)

Figure* 4.2. *a)-b) Change of categoricity in HIR. The dotted line shows categoricity in HIR. It can be seen nicely how it flattens out as the space radius is increased.



(c)
uncertainty = 3.254 (VSM)
uncertainty = 3.318 (HIR, **radius = max d_E + 1)**

(d)
uncertainty = 3.254 (VSM)
uncertainty = 3.322 (HIR, radius = **max d_E + 100**)

Figure* 4.2. *c)-d) Change of categoricity in HIR. The dotted line shows categoricity in HIR. It can be seen nicely how it flattens out as the space radius is increased.

In Figure 4.2 the X-axis denotes the documents (from the "7.b" document to the number "12.a" document) and Y-axis the similarity values. The solid line represents the categoricity of the VSM, which is unchanged in all cases (*a*-*d*). The dotted line shows the extent of categoricity in HIR model. The radius rises from case (*a*) to case (*d*). It can be seen clearly that the line is the flattest in the case (*d*), which means that the system is uncertain, i.e., its categoricity is the lowest. It demonstrates that the system becomes more uncertain by increasing the radius.

4.8 Modifiable Categoricity at Lower Re-Computation Costs by Using HIR [THESIS 2.d]

Uncertainty appears as a degree to which retrieval system is categorical in its answers. In the traditional VSM model, it depends on the similarity measure and/or the weighting scheme. It is shown, in [P1], that modifiable categoricity can be obtained at much lower re-computation costs using HIR model.

Table 4.1 shows the categoricity of four VSM similarity measure and three weighting scheme. Based on this table, it can be clearly seen that:

(i) the usage of the Cosine/Dice similarity measures resulted in the least entropy reduction for every weighting scheme;

(ii) the usage of the *tfn* (normalised frequency) weighting scheme reduces entropy to the greatest extent for every similarity measure.

So far, Cosine measure with *tfn* weighting scheme — one of the most commonly used VSM measures in practice — is the less categorical in its answers. Additionally, it can be seen from table 5.1:

(iii) that the amount of information *U* does not depend on the weighting scheme using Cosine measure,

(iv) and the amount of information *U* only slightly depends on the similarity measure using *tfn* weighting scheme.

Therefore, it is not enough to change only the weighting scheme or the similarity measure, but both of them need to vary to obtain a more categorical system. This in turn yields costly re-computation of both weights and similarity measure values; and the same answers set containing the same document with the same order cannot be guaranteed, because the similarity measures do not preserve the rank order.

4.8.1 Computational Complexities to Obtain Different Categoricities

It was shown in chapter 4.5 that in HIR the categoricity could be varied without taking another weighting scheme and without changing the similarity measure: it will suffice to vary the radius of the space in order to change the categoricity of answers. Due to this property, HIR represents the advantage of being a means to make the categoricity of the Cosine- and *tfn*-based VSM adjustable depending on only one control variable, namely the radius of the C-KHS. Thus, a modifiable categoricity can be obtained at much lower re-computation costs: only the similarity values need to be re-computed but not the weights. In addition, the rank order and the same answer set could be guaranteed, because the HIR and Cosine-based VSM are equivalent using *tfn* weighting scheme.

Let us assume that the user wants to retrieve the same answers set, i.e., same documents, same ranking, but having different categoricities. It is shown that for this purpose it is more advantageous to use HIR than the traditional Cosine measure with *tfn* weighting scheme.

Given a finite set D of elements called *documents*:

D_j, , $j = 1, \ldots, m \in$ **N** (**N** denotes the set of natural numbers)

and a finite set T of elements called index *terms*:

t_i, $i = 1, \ldots, n \in$ **N** (**N** denotes the set of natural numbers)

and a query Q.

It will be estimated the computational complexities to obtain different categoricities:

- In HIR, based on formula of the Hyperbolic similarity measure *Hyp* (it was introduced in chapter 3.3) is the following:

  $$Hyp = \left(\ln \left(e \cdot \frac{r + \sqrt{\sum_{i=1}^{n} \left(w_{ij} - q_i\right)^2}}{r - \sqrt{\sum_{i=1}^{n} \left(w_{ij} - q_i\right)^2}} \right) \right)^{-1}$$, in other form:

  $$Hyp = \frac{1}{1 + \ln \frac{r + A}{r - A}}, \text{ where } A = \sqrt{\sum_{i=1}^{n} \left(w_{ij} - q_i\right)^2}$$

  So far, to obtain a different categoricity system only the similarity values *Hyp* need to be computed *m* times, without the computation of *A*, because only the radius changes and the weighting scheme is unvaried.

  So the computational complexity is:

  $mK = O(m)$,

  where *K* is a constant corresponding to the time required by additions and multiplications. It can be clearly seen that it simply depends on the number of documents.

- In VSM — based on Cosine measure with *tfn* ¾ it is not enough to change only the weighting scheme or the similarity measure, but both of them need to vary to obtain a more categorical system. So *maxNorm* scheme, with the simplest similarity measure: Dot product was selected, because *maxNorm* scheme yields to the most categorical system.

$$Dot = \sum_{i=1}^{n} w_{ij} \cdot q_i \text{ , and } maxNorm\ w_{ij} = \frac{f_{ij}}{\max f_{kj}}$$

So far, to obtain a more categorical system the Dot measure needs to be computed $n \times m$ times, and the weights also $n \times m$ times, because the weighting scheme also changes.

So the computational complexity is:

$2nm + C = O(mn)$,

where C is a constant corresponding to the time required by additions and multiplications. It can be clearly seen that it depends on both the number of documents and the number of index terms.

It follows that using HIR for changing the categoricity is faster, i.e.,

$2n + \alpha = O(n)$ (where α is a constant),

than using VSM for this purpose. Thus, a modifiable categoricity can be obtained at lower re-computation costs: only the similarity values need to be re-computed but not the weights. Additionally, the same answers set containing the same document with the same order could be guaranteed.

4.8.2 Experimental Results

Experimental results illustrate also that using HIR for changing the categoricity is faster, than using VSM. For these purpose three *term-by-document* matrices were used. The first matrix includes about the same number of documents and index terms. The second matrix includes much more index terms, than documents. The third matrix includes much more documents, than index terms. The matrices are the following:

(i) It derives from the ***Belief Database*** (section 2.2.3) containing 2704 belief text and 2607 index terms (n=m). The average number of terms per text was 15.

(ii) It derives from a part of ***Reuters Database*** (section 2.2.2) containing 7000 documents and 32589 index terms (n>>m). The average number of terms per document was 73.

(iii) It is a ***simulated*** — not related to a real database — ***term-by-document*** matrix to examine the influence of a database containing much more documents than index terms. For this purpose a *term-by-document matrix* was generated using a program written in C language. The dimension of the matrix is 2000 x 100000 simulating 100000 documents, and 2000 index terms (n<<m). The average number of terms per document is 15, which was generated by RAND (random number variable) function.

A computer program written in the C language was implemented to determine the computational time of the weights (*tfn*, *maxNorm*) and similarity values (Cosine, Hyperbolic, Dot). The computational time was measured by using the *clock* function. The time is given in seconds and one clock tick is 1 ms.

The computational time of the weights — for three different weighting scheme—, and similarity values — for three different similarity measures —, and its sums can be seen in table 4.4 (for Belief Test Database), in Table 4.5 (for Reuters Test Collection) and in table 4.6 (for the simulated *term-by-document matrix).*

Table 4.4. Running time (in seconds) of the weights and similarity values computation for Belief Database containing 2704 documents, and 2607 index terms (on AMD ATHLON 1.9 GHz, 512 MB RAM computer)

Computational time of **weights**		***Computational time of*** **similarity values**		
tfn	*maxNorm*	Cosine	Dot	Hyperbolic
0.5548 s	0.539 s	1.2354 s	1.2354 s	1.3064 s

Computational time of **weights + similarity values**	
tfn + Cosine	1.7902 s
maxNorm + Dot	1.7744 s
tfn + Hyperbolic	1.8612 s

Table 4.5. Running time (in seconds) of the weights and similarity values computation for Reuters Database containing 7000 documents and 32589 index terms (on AMD ATHLON 1.9 GHz, 512 MB RAM computer)

Computational time of **weights**		***Computational time of*** **similarity values**		
tfn	*maxNorm*	Cosine	Dot	Hyperbolic
6231.4 s	4684.4 s	2102.7 s	2109.9 s	2294.5 s

Computational time of **weights + similarity values**	
tfn + Cosine	8334.1 s
maxNorm + Dot	6794.3 s
tfn + Hyperbolic	8525.9 s

Table 4.6. Running time (in seconds) of the weights and similarity values computation for simulated *term-by-document* matrix containing 100000 documents and 2000 index terms (on AMD ATHLON 1.9 GHz, 512 MB RAM computer)

Computational time of **weights**		***Computational time of*** **similarity values**		
tfn	*maxNorm*	Cosine	Dot	Hyperbolic
3299.3814 s	573.972 s	377.1814 s	273.1468 s	362.1048 s

Computational time of **weights + similarity values**	
tfn + Cosine	3676.5628 s
maxNorm + Dot	847.1188 s
tfn + Hyperbolic	3661.4862 s

The result of the comparison (running time of varying uncertainty for the same ranking order) can be seen in Table 4.7.

Table 4.7. Running time (in seconds) of computing the similarity values to obtain different categoricity in HIR, and VSM.
(on AMD ATHLON 1.9 GHz, 512 MB RAM computer)

	for Belief Database	***for Reuters Database***	***for simulated term-by-document matrix***
Hyperbolic measure	1.8612	8525.9	3661.5
VSM measure (from Cosine with *tfn* to Dot with *maxNorm*)	3.5646 (1.7902+1.7744)	15128.4 (8334.1+ 6794.3)	4523.8 (3676.6+847.2)

Table 4.7 shows that the time of modifying categoricity in HIR is much less than in the VSM (from Cosine with *tfn* to Dot with *maxNorm*). In HIR to obtain different categoricity, only the radius of the space must be changed (e.g., from $r = \max d_E(QD) + 30$ to $r = \max d_E(QD) + 10^{-15}$), so only a part of the similarity measure needs to be re-computed. In VSM the weighting scheme and similarity measures are both changed — from Cosine measure with *tfn* weighting scheme to Dot product with *maxNorm* weighting scheme —, so they both need to be re-computed. It can be seen that it is more advantageous to use HIR, than traditional Cosine measure with *tfn* weighting scheme; Cosine measure is replaceable with HIR using *tfn,* because they are equivalent to each other (it was proved in chapter 3.4.1).

Table 4.8. Time reduction obtained by using HIR for a modifiable categoricity system.
(on AMD ATHLON 1.9 GHz, 512 MB RAM computer)

	for Belief Database	***for Reuters Database***	***for simulated term-by-document matrix***
Running time in HIR	52 %	56.35 %	80.9 %
Running time in VSM	100 %	100 %	100 %
Acceleration	**48** %	**44** %	**19** %

Table 4.8 illustrates that using HIR for a system with modifiable categoricity a time reduction of 20-50 % can be obtained. Consequently, a system with modifiable categoricity — with fixed answers set and rank order — can be obtained at much lower re-computation costs by just changing the radius in HIR.

CHAPTER 5

EFFICIENT WAY TO COMPUTE TERM DISCRIMINATION VALUES

In this chapter the concept of UDO (Uncertainty Decreasing Operation) is proposed as a theoretical background for term discrimination power, and it is applied to the computation of term discrimination values. It is shown that the UDO-based computation is faster and its application is not restricted to the Vector Space Model.

5.1 Motivation

The concepts of Shannon information and entropy have been applied to a number of information retrieval tasks such as to formalise the probabilistic model, to design practical retrieval systems, to cluster documents (Fujii and Ishikawa, 2001), and to model texture in image retrieval (Yoo et al., 2002).

In my dissertation, the concept of entropy is used for a different purpose. It was shown in theorem 4.2, that any positive RSV-based (Retrieval Status Value) retrieval system may be conceived as a special probability space in which the amount of the associated Shannon information is being reduced; in this view, the retrieval system is referred to as UDO (Uncertainty Decreasing Operation). The concept of UDO will be proposed as a theoretical background for term discrimination power, and it will be applied to the computation of term discrimination values in any RSV-based retrieval model. The UDO-based computation, however, presents advantages over the vector-based calculation: it is faster, easier to assess and handle in practice and its application is not restricted to the Vector Space Model.

5.2 Term Discrimination Model

The Term Discrimination Model (TDM) was introduced in (Salton, Yang, and Yu, 1974; Salton, Yang, and Yu, 1975) as a contribution to the automatic indexing theory in the Vector Space Model of information retrieval. The TDM is based on the underlying assumption that a "good" term causes the greatest possible separation of documents in the vector space, whereas a "poor" term makes it difficult to distinguish one document from another. Each term under focus is assigned a Term Discrimination Value (TDV) defined as the difference between space "densities" before and after removing that term. The space "density’" Δ is defined as the average pairwise similarity s between documents D_j ($j = 1\ldots m$):

$$\Delta = \frac{1}{m(m-1)} \sum_{\substack{j,k=1 \\ j\neq k}}^{m} s\,(D_j, D_k) \tag{5.1}$$

Alternatively, a space density Δ can be computed — faster — as the average similarity between documents and a centroid document (defined as one in which the terms have frequencies equal to their average frequencies across the collection of documents). Let Δ_{bi} and Δ_{ai} denote the space "densities" before and after removing term t_i respectively; then the TDV_i of term t_i is defined as follows:

$$\text{TDV}_i = \Delta_{bi} - \Delta_{ai} \tag{5.2}$$

The best discriminators generally have positive TDVs, whereas the worst discriminators usually have negative TDVs. Terms having TDVs around zero do not modify the space density considerably when used as index terms. The TDV can be used for the following purposes:

(i) To decide which terms should be used as index terms (Yu, and Salton, 1977): terms with average document frequencies (between approximately m/100 and m/10) usually have positive TDVs, and can be used directly for indexing purposes; terms whose document frequency is too high generally have negative TDVs, and are worst discriminators; too rare or specific terms have TDVs near zero, and should not be used directly as index terms.

(ii) Weights computation for terms (Salton, 1986): while the Inverse Document Frequency (IDF) method prefers low frequency terms, they are not preferred in the TDM; thus, in the TDM, the weight w_{ij} of term t_i in document D_j should be computed as $w_{ij} = f_{ij}\cdot\text{TDV}_i$ (instead of $w_{ij} = f_{ij}\cdot\text{IDF}_i$).

(iii) Thesaurus construction (Crouch, and Yang, 1992): TDVs of terms are used to construct thesaurus classes.

Practical research carried out in TDM has since highlighted several insights as follows. Crouch and Yang (1992) and Dubin (1995) have found that there is no direct and exact correlation between discrimination value on the one hand, and documents and term frequency on the hand. The computation of the TDV (formula 5.2) is expensive, which hinders its practical application; Willett (1985) developed a faster method, he also showed that whether a term is a poor or good discriminator depends on the similarity measure used: the dot product yields a monotonically decreasing relationship between TDV and document frequency, the Euclidean distance leads to an increasing relationship. As shown also in (Dubin, 1995) the TDV depends not only on the similarity measure used but also on the weighting method and stop list used. In light of these results, it turns out that there are different correlations between TDV, weighting schemes and similarity measures.

The TDM is based on topological concepts such as: separation, distinguishable, density and sparsity. This view is being applied to the Vector Space Model whose typical space is the n-dimensional orthonormal linear space, i.e., each dimension corresponds to a term, the documents are represented as vectors of weights, the fact that the terms are considered to be independent of one another is modelled by the pairwise perpendicular coordinate axes, the usual similarity measures (Dot product, Cosine measure, Dice's coefficient, Euclidean distance) make sense because the space is Euclidean. Thus, the TDM as a theory and the TDV as a computation method cannot be applied to other — than the Vector Space Model — RSV-based information retrieval model which does not use linear or Euclidean space, and hence whose similarity is not based on inner — or dot — product. However, both from a practical and theoretical point of view, it would be useful to have a means to compute terms TDV in any RSV-based retrieval model.

5.3 Experiment: UDO as an Entropy-Based Theory for Term Discrimination and its Computation [THESIS 3.a]

The concept of UDO introduced in chapter 4.2 is valid for any RSV-based retrieval model; it does not necessarily assume the existence of a linear space. It is proposed in [P3] that the UDO view of RSV-based retrieval models be applied to the study and computation of "discriminatory" or "separation" power of individual terms.

Let Q_i denote a single term query containing exactly one term, t_i, where the term t_i is selected from a list of index terms. Then the corresponding entropy H_i (computed using formula 5.1) will be a measure of the extent to which the term t_k is able to reduce the retrieval system's uncertainty in selecting documents (for returning answers). Thus, in the UDO view, the TDV of a term is based on how much it reduces this entropy — associated to a probability space — rather than how much it reduces space density in Euclidean — and hence topological — space.

In what follows, experimental evidence will be shown using the ADI test collection described in section 2.2.1, for the computation of UDO-based TDV, and how it compares to vector-based TDV. Table 5.1 shows the ADI statistics — containing 82 homogeneous articles from the information science with 915 index terms — for the experiments. The terms were selected automatically, they were TIME stoplisted, and Porter stemmed.

Table 5.1. Statistics for the ADI test collection used in the UDO-based TDV computation.

Subject Area	Information Science
Type	Homogeneous
No. of Documents	82
No. of Terms	915

The experiment was performed using Mathcad in the following form:

given a finite set D of elements called *documents*:

D_j, , $j = 1, \dots, 82 \in$ **N** (**N** denotes the set of natural numbers),

and a finite set T of elements called index *terms*:

t_i, $i = 1, \dots, 915 \in$ **N** (**N** denotes the set of natural numbers),

every document D_j is assigned a vector:

$\mathbf{w}_j = (w_{ij})_{i=1,\dots,915}$ of *weights*,

where $w_{ij} \in$ **R** (**R** denotes the set of real numbers) denotes the *weight* of term t_i for document D_j.

The similarity measure used was the Cosine measure:

$$\sigma\,(\mathbf{w}_j, \mathbf{q}) = \frac{\sum_{i=1}^{n} w_{ij} q_i}{\sqrt{\sum_{i=1}^{n} w_{ij}^2 \cdot \sum_{i=1}^{n} q_i^2}},$$

normalised inverse document frequency scheme was used to compute the weights:

$$w_{ij} = \frac{f_{ij} \cdot \log \frac{m}{F_i}}{\sqrt{\sum_{i=1}^{n} \left(f_{ij} \cdot \log \frac{m}{F_i} \right)^2}},$$

where, f_{ij} denotes the number of occurrences of term t_i in document D_j,

F_i is the number of documents in which the term t_i occurs.

In the experiment the term discrimination values are computed in the following ways:

Case 1.: UDO-based term discrimination values are computed as entropy reductions, i.e.,:

$H_{max} - H$ (in %)

where, H_{max} =log$_2$ m

$$H = -\sum_{j=1}^{m} p_j \cdot \log_2 p_j ,$$

and probabilities p_j are in the following form:

$$p_j = \frac{r_j}{\sum_{k=1}^{m} r_k} j = 1,\ldots, m$$

r_j denotes the RSV (retrieval status value) of document D_j relative to query Q.

The UDO-based TDVs of terms can be seen in figure 5.1. E.g., the 150th term, reduces entropy by 30%.

Case 2.: The vector space term discrimination values are calculated as space density variation (formula 6.2), i.e.:

$$\text{TDV}_i = \Delta_{bi} - \Delta_{ai}$$

where Δ_{bi} and Δ_{ai} denote the space 'densities' before and after removing term t_i respectively.

The space 'density' Δ is defined as the average pairwise similarity s between documents D_i ($i = 1\ldots82$):

$$\Delta = \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{m} s\,(D_i, D_j) .$$

Figure 5.2 shows the vector-based TDVs calculated as space density variations using centroid document. E.g. the 150th term has negative values meaning that it is a poor discriminator.

It can be seen from figure 5.1 a-c that about half the terms have their UDO-based TDV equal to 100%, and figure 5.3 a-c shows that the frequency of each such term is equal to 1, i.e., each such term occurs in exactly one document. The vector-based TDV of every such term is positive with a mean value of 6.4×10^{-6}. These terms are good discriminators. This result can be seen in figure 5.2 a-c, and figure 5.4.

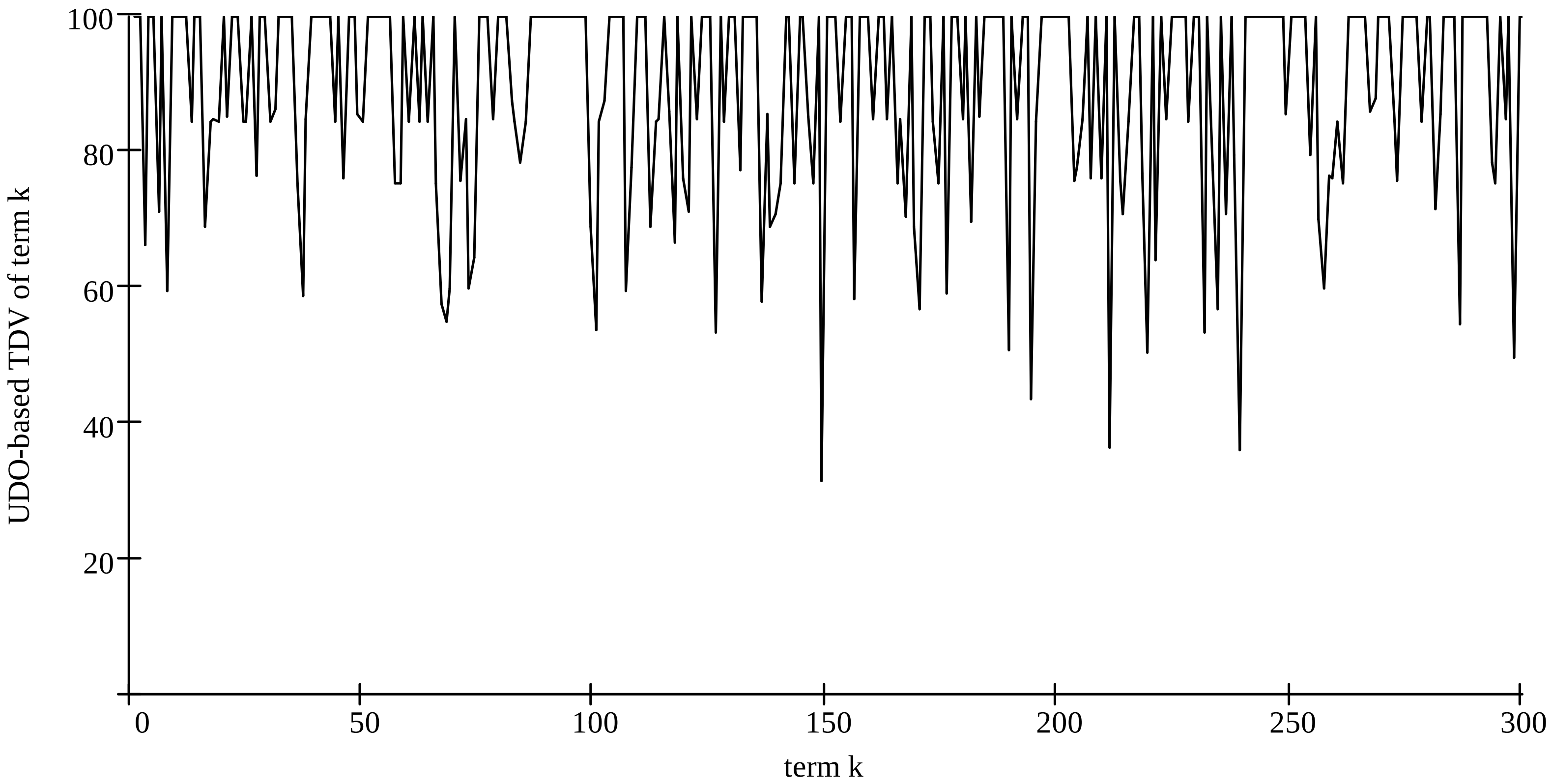


Figure 5.1. a) UDO-based TDV for the ADI test collection for the index terms (1-300).
Discrimination values of terms are computed as entropy reductions. On the horizontal axis, term k means term t_k (the kth term in the list of index terms). On the vertical axis, the corresponding TDV is shown as the percentage of entropy reduction from the maximum entropy value. The Cosine similarity measure, and the ***n-idf*** weighting scheme were used. E.g., the 150th term, reduces entropy by 30%.

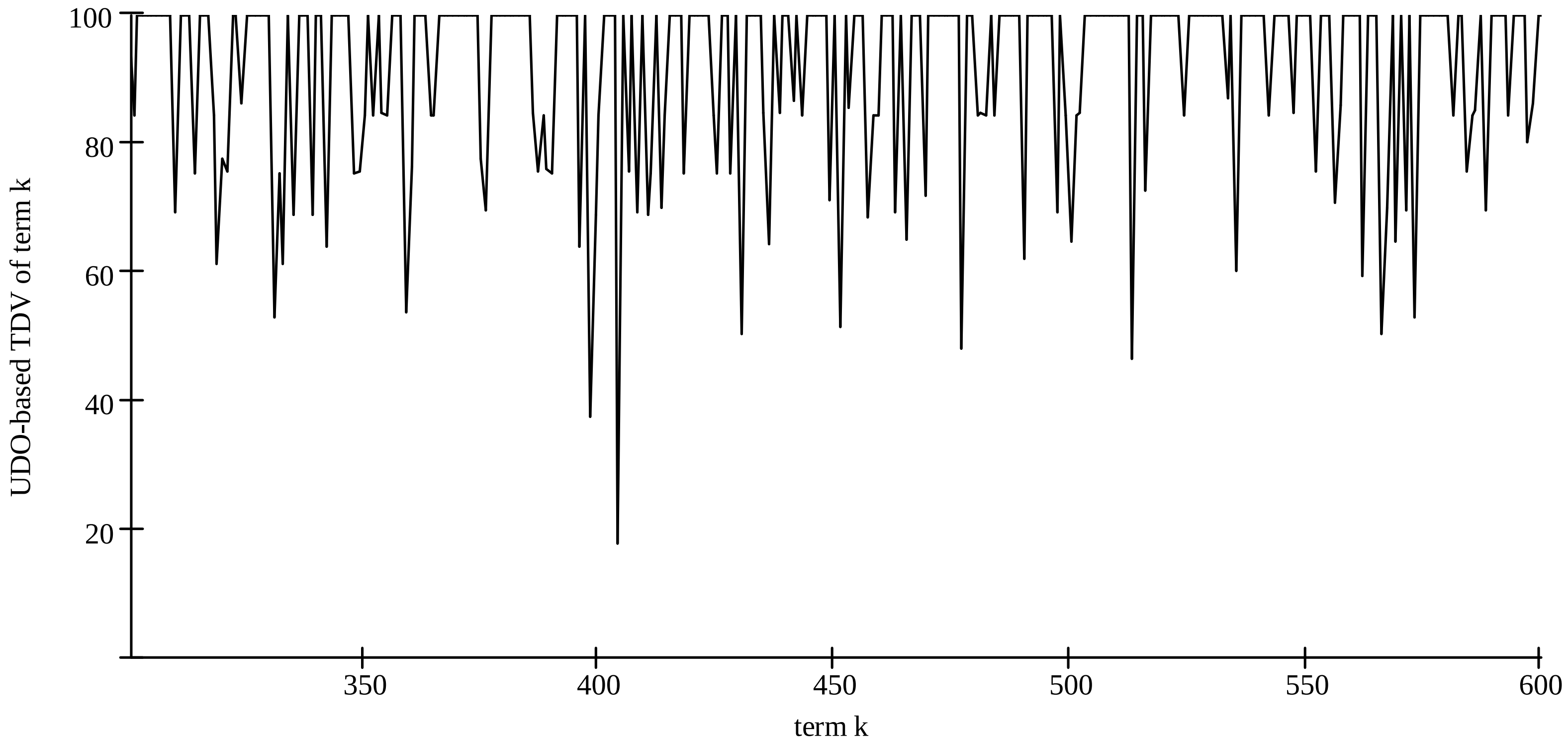


Figure 5.1. b) UDO-based TDV for the ADI test collection for the index terms (301-600).
Discrimination values of terms are computed as entropy reductions. On the horizontal axis, term k means term t_k (the kth term in the list of index terms). On the vertical axis, the corresponding TDV is shown as the percentage of entropy reduction from the maximum entropy value. The Cosine similarity measure, and the ***n-idf*** weighting scheme were used.

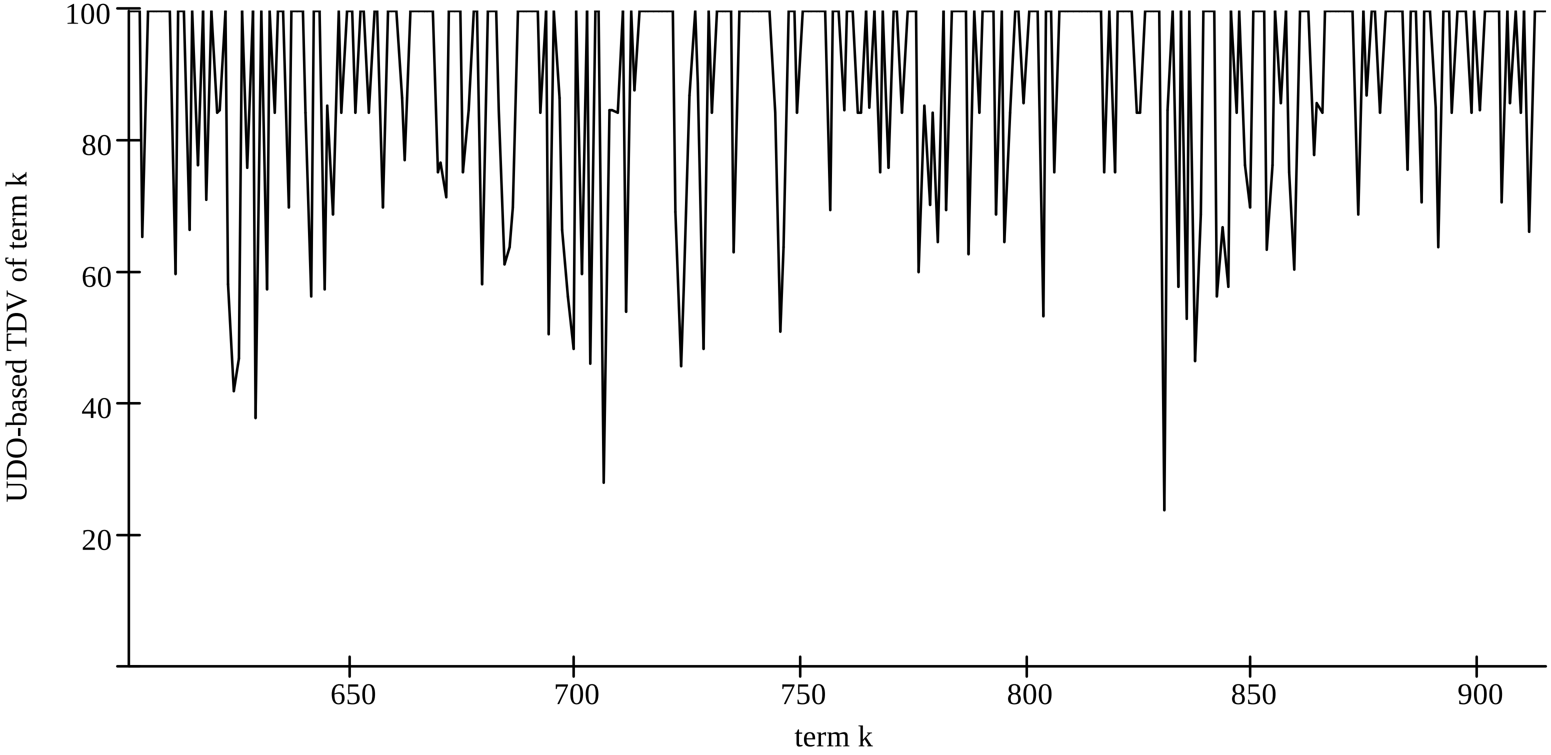


Figure 5.1. c) UDO-based TDV for the ADI test collection for the index terms (601-915).
Discrimination values of terms are computed as entropy reductions. On the horizontal axis, term k means term t_k (the kth term in the list of index terms). On the vertical axis, the corresponding TDV is shown as the percentage of entropy reduction from the maximum entropy value. The Cosine similarity measure, and the ***n-idf*** weighting scheme were used.

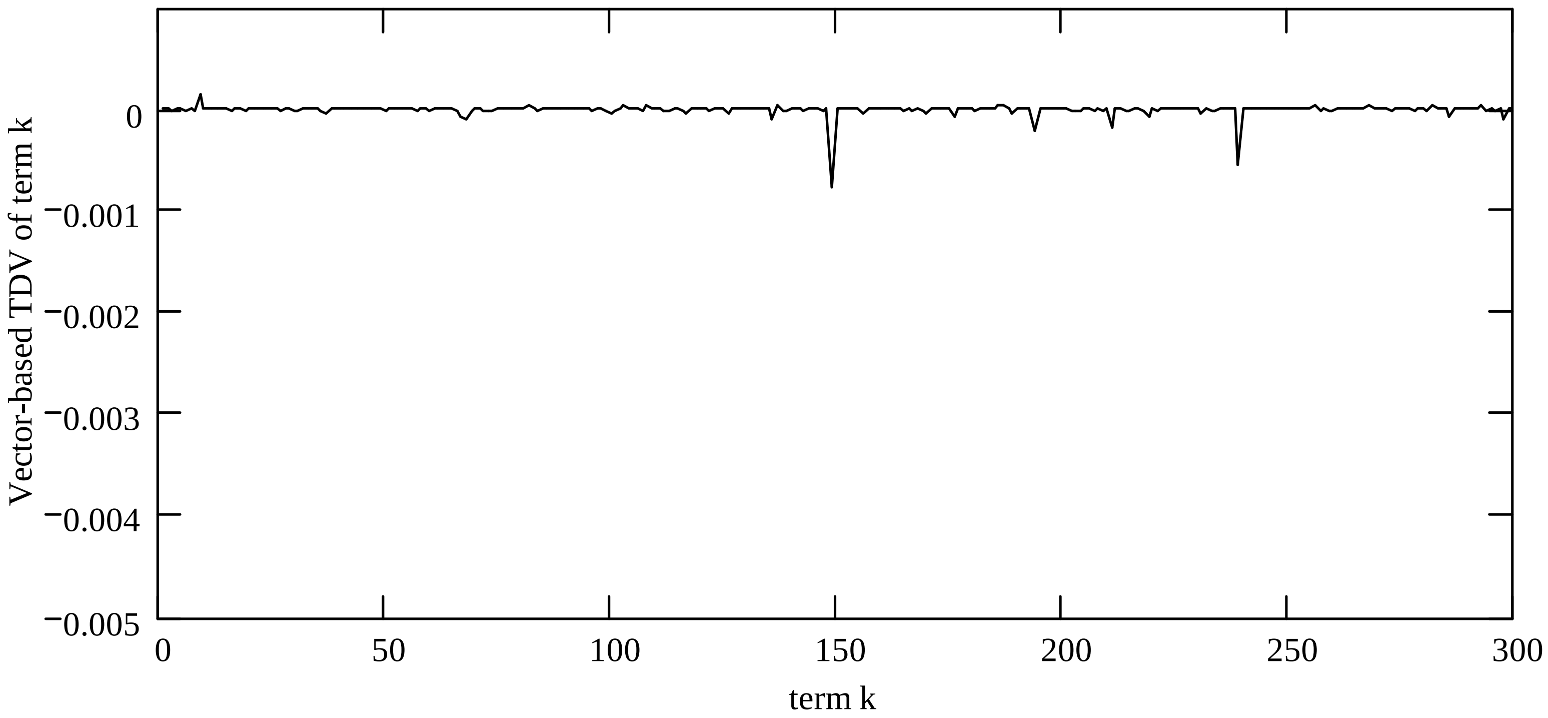


Figure 5.2. a) Vector-based TDV of the terms (1-300) for the ADI test collection.
On the horizontal axis, term *k* means term t_k, i.e., the *k*th term in the list of index terms.
On the vertical axis, the corresponding TDV is shown computed using centroid document.
The Cosine similarity measure and the normalised inverse document frequency weighting scheme were used

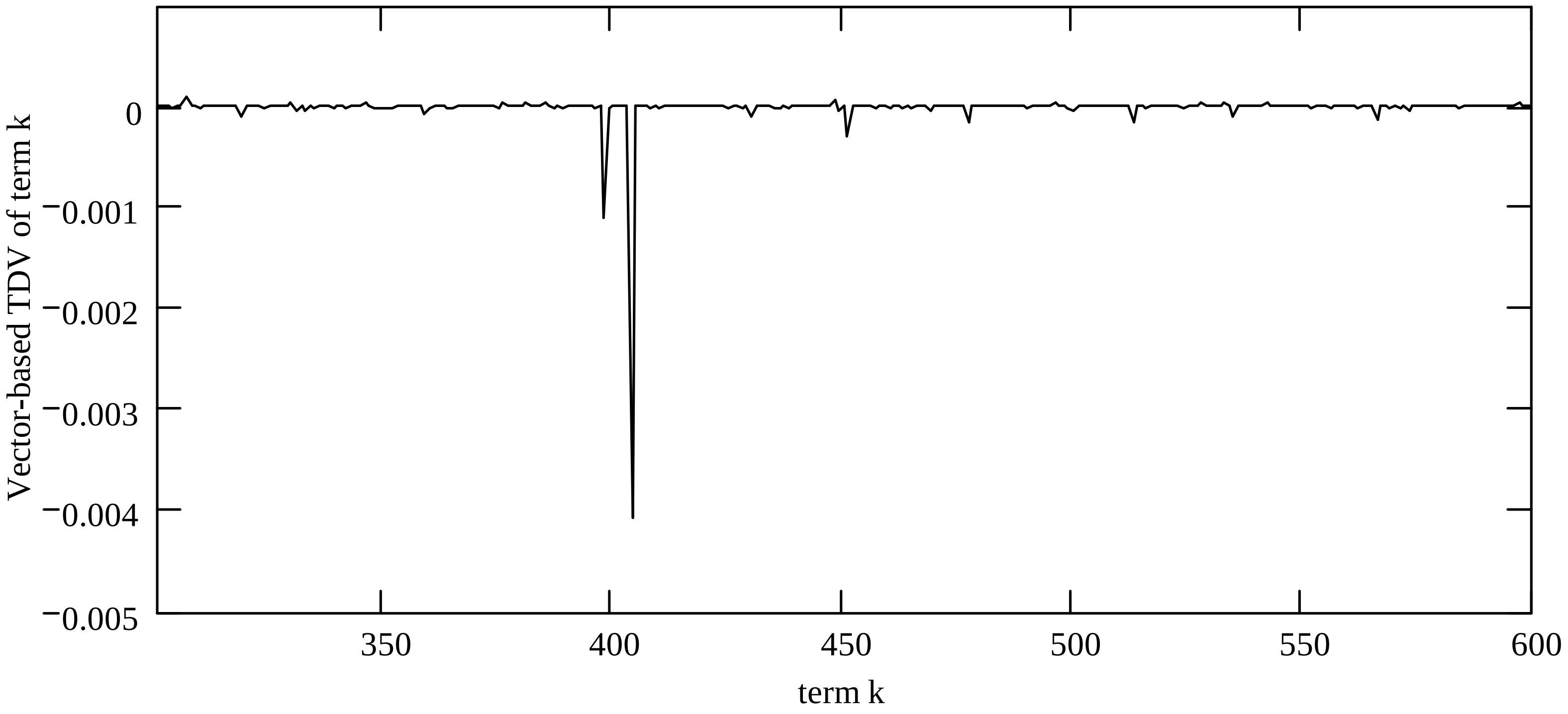


Figure 5.2. b) Vector-based TDV of the terms (301-600) for the ADI test collection.
On the horizontal axis, term *k* means term t_k, i.e., the *k*th term in the list of index terms.
On the vertical axis, the corresponding TDV is shown computed using centroid document.
The Cosine similarity measure and the normalised inverse document frequency weighting scheme were used

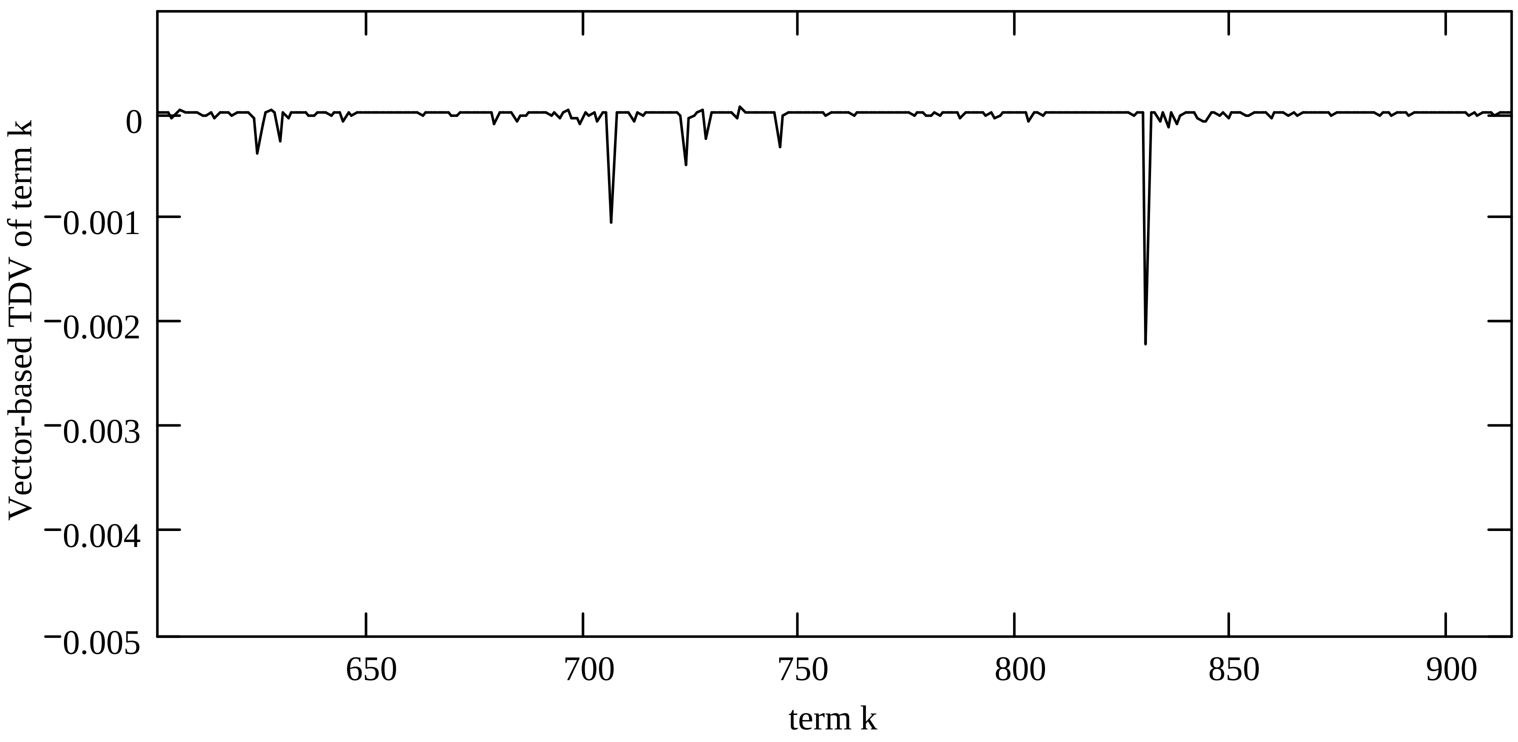


Figure 5.2. c) Vector-based TDV of the terms (601-915) for the ADI test collection.
On the horizontal axis, term ***k*** means term ***t_k***, i.e., the ***k***th term in the list of index terms.
On the vertical axis, the corresponding TDV is shown computed using centroid document.
The Cosine similarity measure and the normalised inverse document frequency weighting scheme were used.

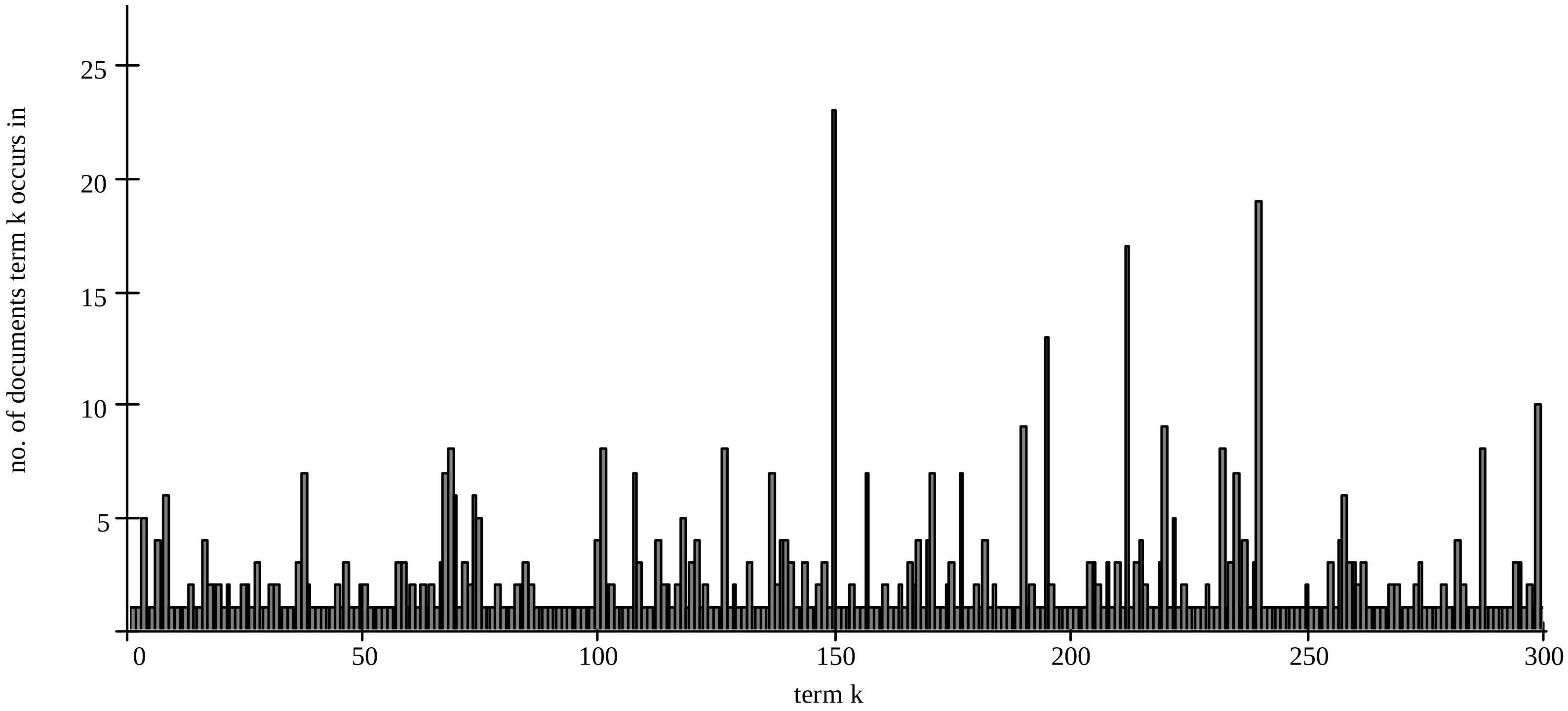


Figure 5.3. a) Document frequency of the terms (1-300) in the ADI test collection.
On the horizontal axis, term ***k*** means the k***th*** term in the list of index terms.
On the vertical axis, the document frequency of term ***k*** is shown, i.e., the number of documents in which the term ***k*** occurs.

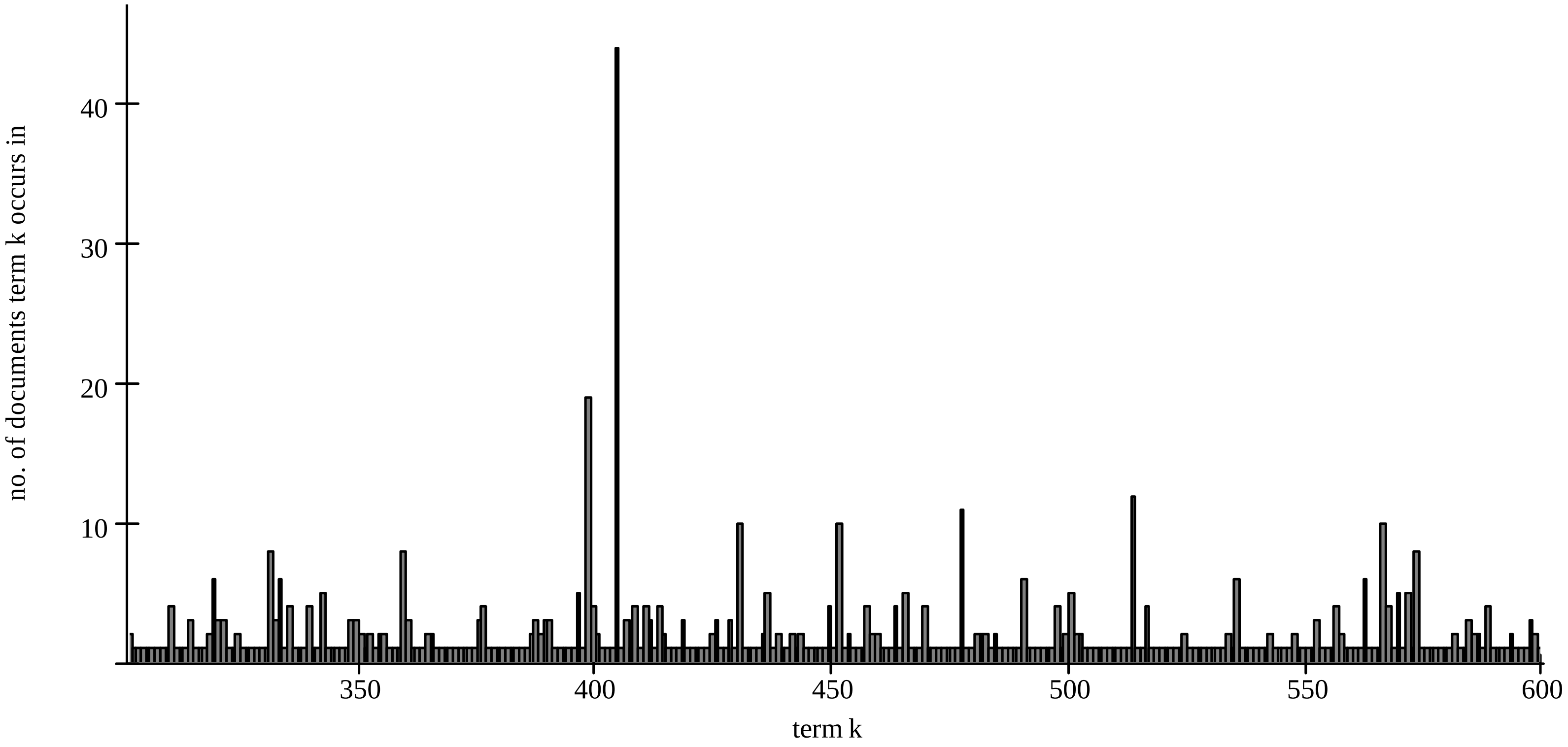


Figure 5.3. b) Document frequency of the terms (301-600) in the ADI test collection.
On the horizontal axis, term ***k*** means the k***th*** term in the list of index terms.
On the vertical axis, the document frequency of term ***k*** is shown, i.e., the number of documents in which the term ***k*** occurs.

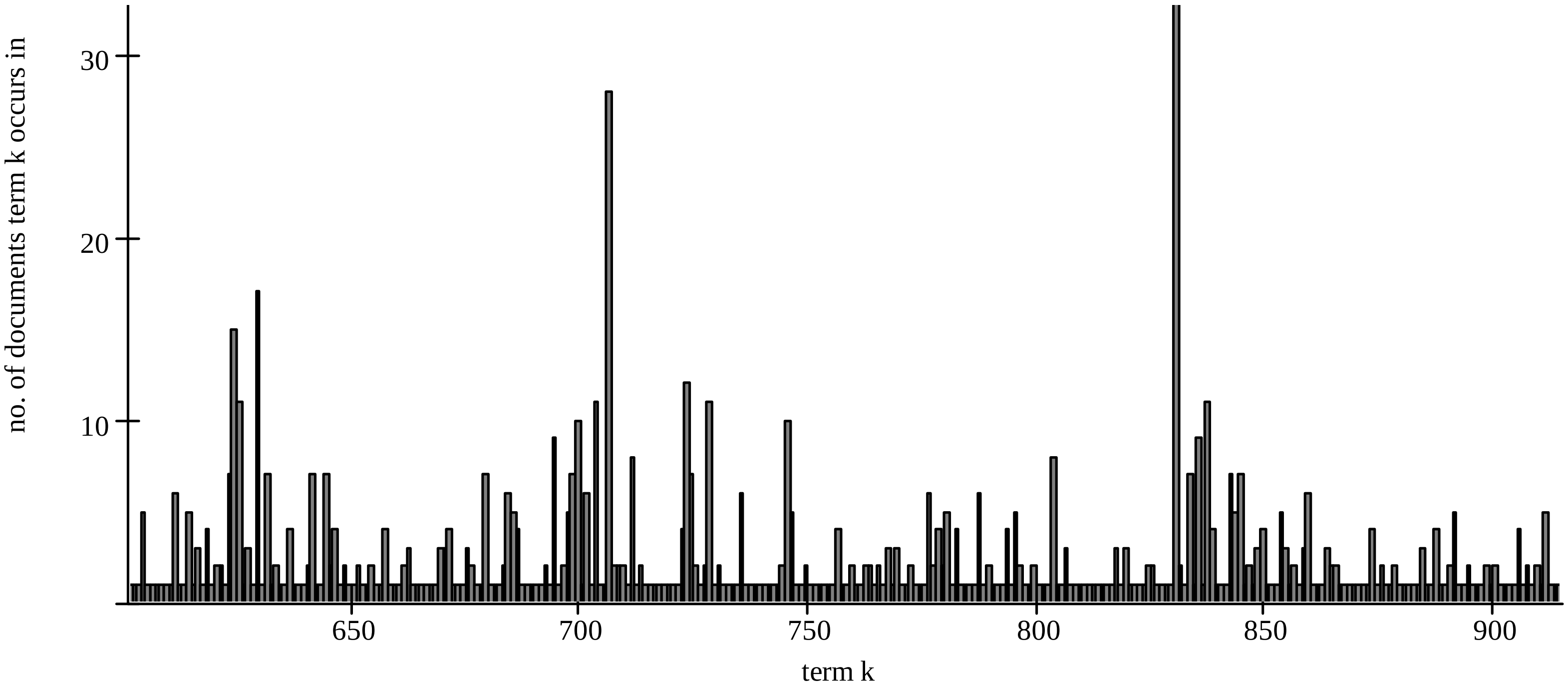


Figure 5.3. c) Document frequency of the terms (601-915) ADI test collection.
On the horizontal axis, term ***k*** means the k***th*** term in the list of index terms.
On the vertical axis, the document frequency of term ***k*** is shown, i.e., the number of documents in which the term ***k*** occurs.

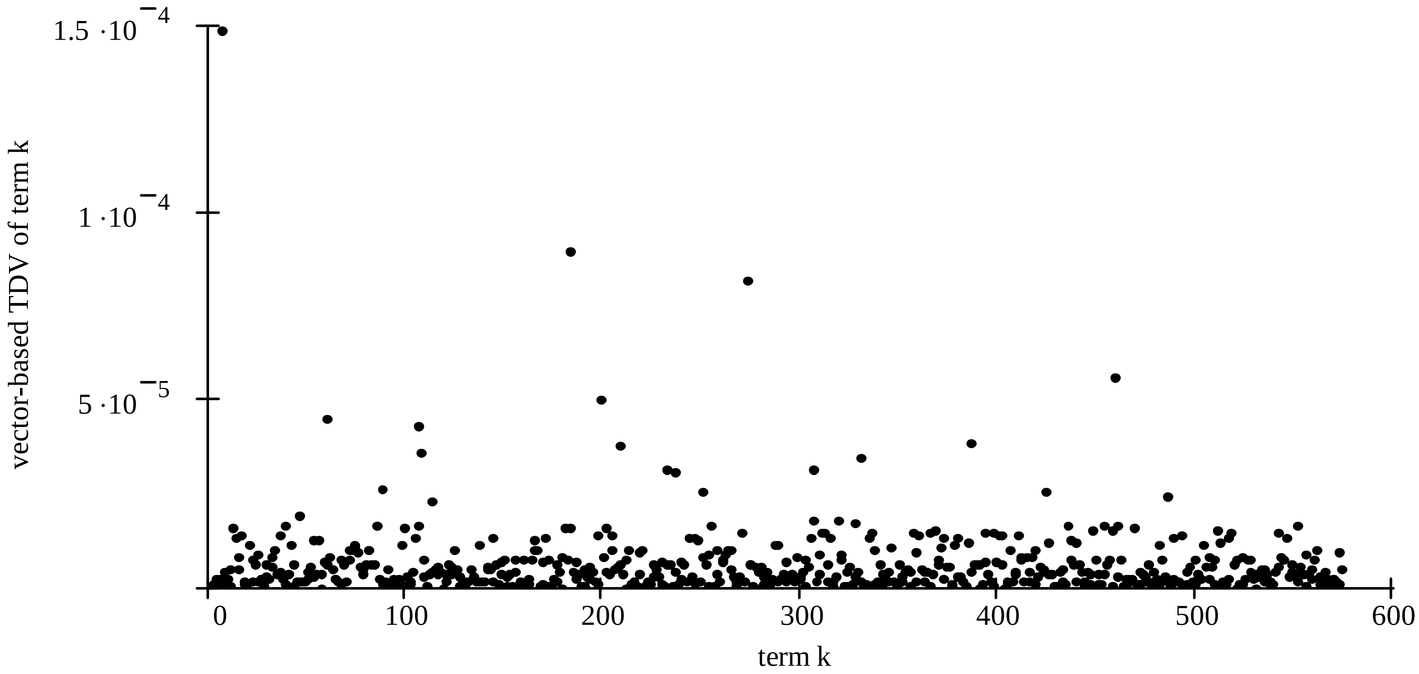


Figure 5.4. Vector-based TDV of those terms whose UDO-based TDV is 100% (ADI test collection). The document frequency of every term is equal to 1. The mean value of TDVs is equal to 6.4×10^{-6} with a standard deviation of 0

Figure 5.5 shows the vector-based TDV of terms whose UDO-based TDV is less than 100% and greater than 80%. The mean of vector-based TDV is equal to 6.056×10^{-6} having zero standard deviation. The document frequency of every such term is 2. Most of these terms have positive vector-based TDV.

Figure 5.6 shows the vector-based TDV of those terms whose UDO-based TDV belongs to the interval [40%; 80%]. The vector-based TDV is practically zero for every such term. Document frequency has a mean equal to 5.066 with standard deviation equal to 2.406.

A few terms reduce entropy with less than 40%, they are poor discriminators. Their document frequency is high, with a mean value of 25 and standard deviation of 8.93. Their vector-based TDV are negative and low values. The dotted line in figure 5.7 shows the document frequencies of terms (scaled for representation and comparison purposes), whereas the solid line shows the vector-based TDVs. There appears to be a symmetry, hence a direct proportionality between document frequency and vector-based TDV in this case.

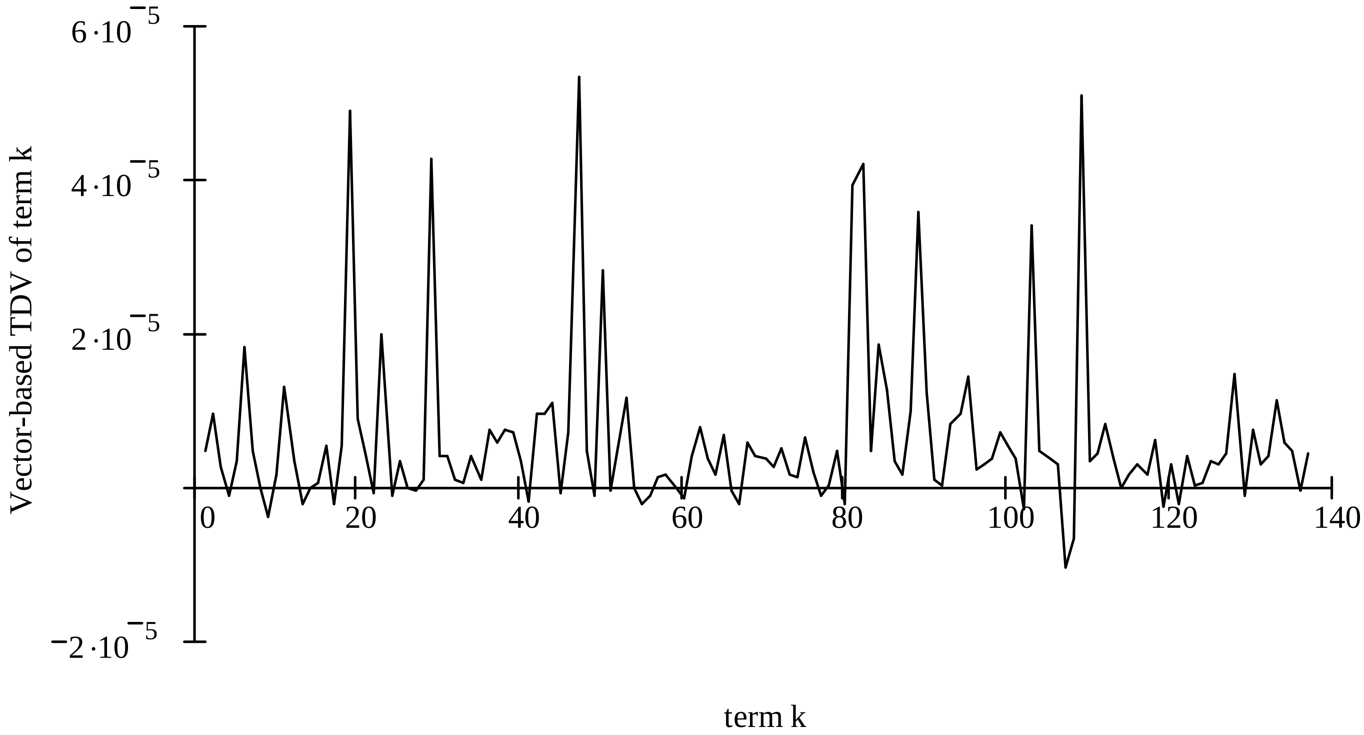


Figure 5.5. The vector-based TDV of terms whose UDO-based TDV is less than 100% and greater than 80%.
The mean of vector-based TDV is equal to 6.056×10^{-6} having zero standard deviation. The document frequency of every such term is 2

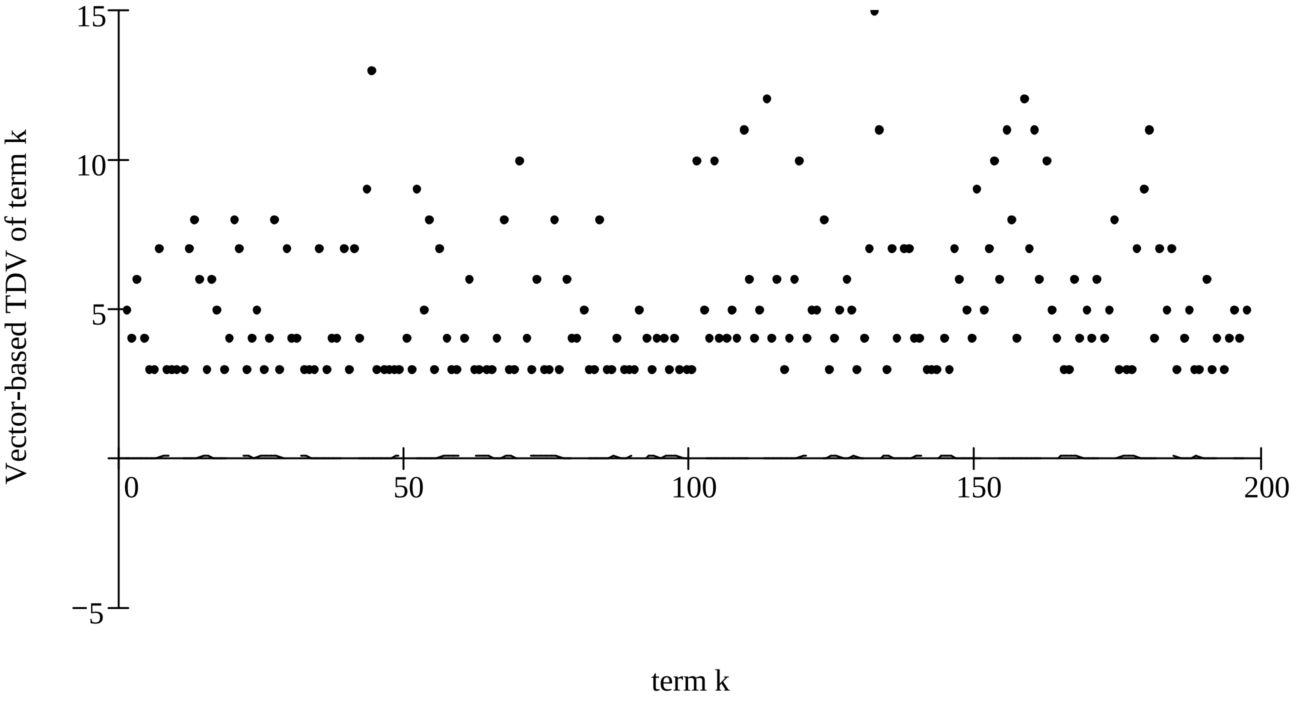


Figure 5.6. The vector-based TDV of those term whose UDO-based TDV belongs to the interval [40%; 80%].

The vector-based TDV (shown as solid line segments along the horizontal axis) is practically zero for every such term. Document frequency (shown as dots) has a mean equal to 5.066 with standard deviation equal to 2.406.

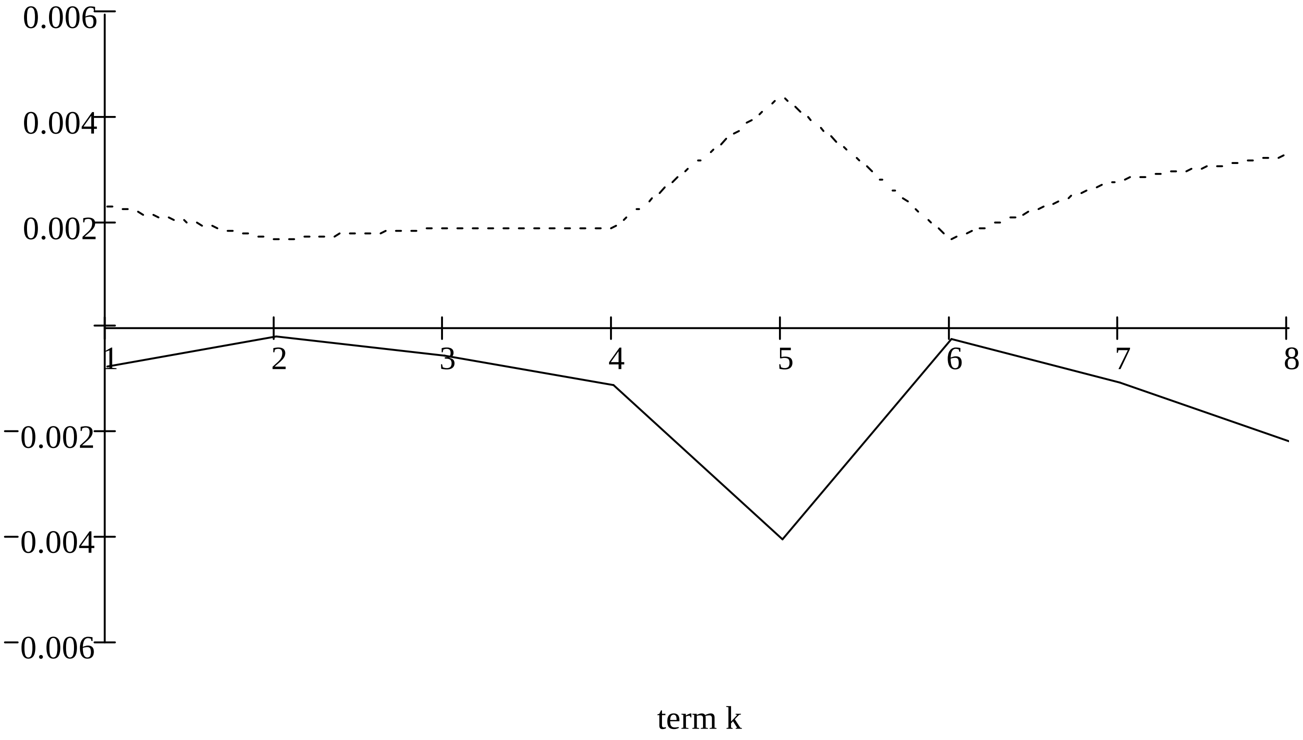


Figure 5.7. Vector-based TDV of those terms whose entropy reduction is below 40% (ADI test collection). TDV is shown as solid line, and the corresponding document frequency as dotted line.

The results of using the UDO-based TDVs method can be summarised as follows:

- the terms that reduce entropy almost entirely, i.e., in the interval (80%; 100%], are very good discriminators and have very low document frequency,
- the terms that hardly reduce entropy, i.e., in the interval [0%; 40%), are poor discriminators and have very high document frequency,
- the terms that reduce entropy in the middle range, in the interval [40%; 80%], are indifferent discriminators and have relatively medium document frequency.

Similar results can be obtained using the vector-based TDV computation, however, the entropy-based TDV calculation method presents the following advantages:

- the UDO-based view does not assume the existence of any linear space like the traditional TDM does;
- the UDO-based view can be applied to compute term TDVs in any RSV-based retrieval model (not just in the Vector Space Model);
- the numeric results (i.e., the entropy reduction values given as %) are easier to assess and handle than the vector-based TDV expressed as fractional real numbers with exponent;
- it is faster than the computation of vector-based TDV with formula 5.2. (more details in chapter 5.4).

5.4 UDO-Based TDV Calculation Method as a Faster Method [THESIS 3.b]

It is shown in [P3] that using the new entropy-based TDV calculation — introduced in section 5.3 — to compute term discrimination values is faster than using the traditional TDM model.

5.4.1 Computational Complexities of the UDO-Based, and Vector-Based Term Discrimination Method

Comparing the UDO-based TDV and the vector-based TDV computation method in the computational complexity aspect, one can state the following:

Given a finite set D of elements called *documents*:

D_j, , $j = 1, \ldots, m \in \mathbf{N}$ ($\mathbf{N}$ denotes the set of natural numbers),

and a finite set T of elements called index *terms*:

t_i, $i = 1, \ldots, n \in \mathbf{N}$ ($\mathbf{N}$ denotes the set of natural numbers).

Case 1.: Term discrimination values are computed as entropy reductions (UDO-based TDV computation method):

Let Q_i denote a single term query containing exactly one term, t_i, where the term t_i is selected from a list of index terms.

Then the corresponding entropy H_i will be a measure of the extent to which the term t_i is able to reduce the retrieval system's uncertainty in selecting documents:

$$H_i = -\sum_{j=1}^{m} p_j \cdot \log_2 p_j \ ,$$

probabilities p_j are of the following form:

$$p_j = \frac{r_j}{\sum_{k=1}^{m} r_k} \ j = 1,\ldots, m$$

r_j denotes the RSV (retrieval status value) of document D_j relative to query Q.

Entropy reduction can be given in the following form:

$$H_{max} - H \text{ (in \%)}$$

where, $H_{max} = \log_2 m$

So far, the similarity measure — to obtain the RSV values r_j — needs to be computed m times; i.e. for all the documents.

So, the complexity of the computation the term discrimination value of a term t_i is:

$$O(m)$$

Case 2.: Term discrimination values are calculated as space density variation (vector-based TDV computation method):

$$\text{TDV}_i = \Delta_{bi} - \Delta_{ai}$$

where Δ_{bi} and Δ_{ai} denote the space 'densities' before and after removing term t_i respectively.

The space 'density' Δ is defined as the average pairwise similarity s between documents D_j:

$$\Delta = \frac{1}{m(m-1)} \sum_{\substack{j,k=1 \\ j \neq k}}^{m} s\,(D_j, D_k)\,.$$

So far, the s similarity measure needs to be computed for all the documents.
So, the complexity of the computation the term discrimination value of a term t_i is:

$$O(m^2).$$

It follows that using the UDO-based method as term discrimination method is faster, i.e., $O(m)$, than using vector-based TDM for this purpose. Thus, the term discrimination values can be computed at lower re-computation costs by using UDO-based method. Additionally, the UDO-based method can be used in any RSV-based retrieval system, not only in the Vector Space Model.

5.4.2 Experimental Results

Comparing the UDO-based TDV and the vector-based TDV computation method in the computational complexity aspect an experiment was performed based on ADI standard test collection (section 2.2.1) using Mathcad.

Given a 82 *documents*:

$$D_j, , j = 1, \ldots, 82 \in \mathbf{N} \text{ (}\mathbf{N}\text{ denotes the set of natural numbers),}$$

and 915 index *terms*:

$$t_i, i = 1, \ldots, 915 \in \mathbf{N} \text{ (}\mathbf{N}\text{ denotes the set of natural numbers).}$$

The similarity measure used was the Cosine measure:

$$s\ (\mathbf{w}_j, \mathbf{q}) = \frac{\sum_{i=1}^{n} w_{ij} q_i}{\sqrt{\sum_{i=1}^{n} w_{ij}^2 \cdot \sum_{i=1}^{n} q_i^2}},$$

and the *tfn* weighting scheme was used to compute the weights:

$$w_{ij} = \frac{f_{ij}}{\sqrt{\sum_{i=1}^{n} f_{ij}^2}}$$

Case 1.: Term discrimination values are computed as entropy reductions (UDO-based TDV computation method):

In section 5.5.1 it was shown that the Cosine similarity measure — to obtain the RSV values r_j — needs to be computed m times; in this case m was *82.*

So the computational complexity is:

$$O(m)$$

Mathcad needs — on AMD ATHLON 1.6 GHz, 512 MB RAM computer — approximately 40s to compute the Cosine similarity measure 82 times.

Case 2.: Term discrimination values are calculated as space density variation (vector-based TDV computation method):

In section 5.5.1 it was shown that the s measure needs to be computed for all the documents, i. e.,

$\frac{m(m-1)}{2}$ times, where $m = 82$; i.e 3321 times.

So the computational complexity is:

$O(m^2)$.

Mathcad needs — on AMD ATHLON 1.6 GHz, 512 MB RAM computer — approximately 1620s to compute the Cosine similarity measure 82 times.

It is shown experimentally also, that using the UDO-based method as term discrimination method is faster, than using vector-based TDM for this purpose. Thus, the term discrimination values can be computed at lower re-computation costs by using UDO-based method.

CHAPTER 6

CONCLUSIONS

The main contributions and the proposed dissertation — both in English and Hungarian — are summarized in the next sections, and then the publications related to this dissertation are listed.

6.1 Theses

1. *Vector Space Model in Non-Euclidean Space*

In general the Euclidean geometry is the only type of space used in the VSM. In information retrieval non-Euclidean spaces are used for information visualization [21][22]. In my dissertation the Vector Space Model was defined over Cayley-Klein Geometry. [**P1**][**P2**][**P6**]

(a) HIR (Hyperbolic Information Retrieval) Model was defined; the similarity measure was derived from the Cayley-Klein hyperbolic distance. [**chapter 3.4**]

(b) It was shown — formally and experimentally —, that the HIR Model is equivalent to the traditional Vector Space Model using a normalized weighting scheme. [**chapter 3.5**]

2. Efficient method to vary retrieval categoricity

The retrieval categoricity of the VSM, and HIR model was investigated and a new efficient way to vary retrieval categoricity was introduced. [**P1**][**P3**]

(a) It was shown that any retrieval model or system based on positive RSV (Retrieval Status Value) may be conceived as a probability space that decreases the amount of the associated Shannon information. [**chapter 4.4**]

(b) It was shown experimentally that the retrieval categoricity of a VSM depends on the similarity measure, and the weighting scheme. Thus, in the VSM the only way to modify the retrieval categoricity is to take a different weighting scheme and/or similarity measure. So far, the Cosine measure with *tfn* weighting scheme — one of the most commonly used — is the least categorical in its answers. Therefore, it is not enough to change only the weighting scheme or the similarity measure, but both of them need to vary to obtain a more categorical system. This in turn yields costly re-computation of both weights and similarity measure values; and the same answers set containing the same document with the same order cannot be guaranteed, because the similarity measures do not preserve the rank order. [**chapter 4.6**]

(c) It was shown experimentally that in HIR the retrieval categoricity depends on the radius of the space. So far, increasing the radius of the hyperbolic space yields a less categorical retrieval system and conversely: decreasing the radius leads to more categorical answers. [**chapter 4.7**]

(d) HIR represents the advantage of being a means to make the categoricity of the Cosine- and *tfn*-based VSM adjustable depending on only one control variable, namely the radius of the space. Thus, a modifiable categoricity can be obtained at much lower re-computation costs: only the similarity values need to be re-computed but not the weights. In addition, the rank order and the same answer set could be guaranteed, because the HIR and Cosine-based VSM are equivalent using *tfn* weighting scheme. [**chapter 4.8**]

3. Efficient method to compute term discrimination values

In information retrieval the documents are represented by index terms created manually or automatically. TDM (Term Discrimination Method) is an automatic method for creating the index terms. In the thesis a new method was developed to the computation of term discrimination values, which presents advantages over the traditional vector-based calculation. It is faster and its application is not restricted to the Vector Space Model. [**P3**]

(a) Based on (2.a) a new method was developed to the computation of term discrimination values, which is not restricted to the Vector Space Model; it can be used in any positive RSV-based information retrieval system. [**chapter 5.3**]

(b) It was shown that the (3.a) method is faster, than the traditional vector-based TDM method. [**chapter 5.4**]

6.2 Tézisek magyar nyelven

Az értekezés új tudományos eredményei az alábbiakban foglalhatók össze:

1. Vektortér modell nem-euklideszi térben

A vektortér modellt hagyományosan euklideszi térben definiálják. Az információ-visszakeresésben nem-euklideszi teret eddig csak vizualizálásra használtak [21][22]. A dolgozatomban megadtam a vektortér modellt Cayley Klein-féle térben. [**P1**][**P2**][**P6**]

(a) Megadtam a hiperbolikus információ-visszakereső (HIR=Hyperbolic Information Retrieval) modellt, ebben a hasonlósági mértéket a Cayley-Klein hiperbolikus távolságból származtattam. [**chapter 3.4**]

(b) Megmutattam — mind formálisan mind kísérleti úton —, hogy a HIR modell ekvivalens a vektortér modellnek a gyakorlatban leginkább használatos változatával. [**chapter 3.5**]

2. *Hatékony módszer megadása a válaszok kategoricitásának változtatására*

A vektortér modellnek nem-euklideszi térben való megadásán túlmenően megvizsgáltam a HIR modell alkalmazását információ-visszakereső rendszer kategoricitási tulajdonságának változtatására, és megadtam egy hatékony módszert a válaszok kategoricitásának változtatására. [**P1**] [**P3**]

(a) Megmutattam, hogy egy pozitív RSV (Retrieval Status Value)-alapú információ-visszakereső rendszerhez bizonytalanság (Shannon-féle információ) rendelhető hozzá, és ez csökken a válaszadás során. [**chapter 4.4**]

(b) Kísérletileg kimutattam, hogy a hagyományos vektortér modell különböző hasonlósági mértékek és különböző súlyszámítási sémák esetében különböző kategoricitású válaszokat ad vissza. Így adott információ-visszakereső rendszer kategoricitásának változtatása a hasonlósági mérték és/vagy a súlyszámítási séma változtatásával valósítható meg. A gyakorlatban leginkább használatos vektortér modell legelterjedtebb hasonlósági mértéke — a Cosine-mérték *tfn* súlyszámítással — rendelkezik a legrosszabb kategoricitási tulajdonsággal, azaz a válaszok itt a legkevésbé kategorikusak. Ahhoz, hogy kategorikusabb rendszert kapjunk, a hasonlósági mérték megváltoztatásán túlmenően egy másik súlyszámítási sémára való áttérés is szükséges; ez magas számítási bonyolultságot jelent, ráadásul ez esetben a rangsortartás és a válaszhalmaz azonossága sem garantált. [**chapter 4.6**]

(c) Kísérletileg kimutattam, hogy a HIR modellben a válaszok kategoricitási tulajdonsága a sugár változtatásával módosul, mégpedig úgy, hogy a sugár növelésével a válaszokra vonatkozó bizonytalanság is nő, ami a kategoricitás csökkenéséhez vezet. [**chapter 4.7**]

(d) Megmutattam, hogy ha változtatható kategoricitású rendszert akarunk megvalósítani, akkor a Cosine-mértékkel *tfn* súlyszámítási sémával rendelkező vektortér modell helyett gazdaságosabb a HIR modell alkalmazása, mert a kategoricitási tulajdonsága csupán a sugár változtatásával módosítható más súlyszámítási séma, illetve más hasonlósági mértékre való áttérés nélkül, és ez alacsonyabb számítási bonyolultságot jelent; ezen túlmenően a rangsor és a válaszhalmaz változatlan marad. [**chapter 4.8**]

3. *Kifejezések diszkriminálási értékeinek hatékony kiszámítása*

Az információ-visszakereső rendszerekben indexkifejezések megállapítására többféle módszer ismert. Ezek közül az egyik automatikus eljárás a TDM (Term Discrimination Method) módszer. A dolgozatomban egy olyan új módszert fejlesztettem ki kifejezések diszkriminálási értékeinek (TDV= Term Discrimination Value) meghatározására, amelynek alkalmazása hatékonyabb információ-visszakereső rendszert eredményez, mint a hagyományos TDM alapú rendszer. [**P3**]

(a) A (2.a)-ban megadott tulajdonságra alapozva kidolgoztam egy olyan módszert kifejezések szétválasztási értékeinek kiszámítására, amely nemcsak vektortér alapú, hanem bármely, RSV-alapú információ-visszakereső rendszerben használható. [**chapter 5.3**]

(b) Megmutattam, hogy a (3.a)-ban megadott módszer hatékonyabb eljárás kifejezések diszkriminálási értékeinek meghatározására, mint a hagyományos vektor-alapú TDM módszer. [**chapter 5.4**]

6.3 Publications and Citations

6.3.1. Publications directly related to the thesis

Papers in international journals:

[P1] **GÓTH, J.,** and SKROP, A. (2005). Varying Retrieval Categoricity Using Hyperbolic Geometry, *Information Retrieval* Vol. 8, no 2, pp. 265-283 (SCI JIF = 1.185), ISSN: 1386-4564. [**thesis 1,2**]

[P2] DOMINICH S., **GÓTH, J.,** KIEZER, T. (2005). NeuRadIR: A Web-Based NeuroRadiological Information Retrieval System. *ERCIM (European Community in Information Technology) News* No 61, pp. 52-53, ISSN: 0926-4981. [**thesis 1**]

[P3] DOMINICH, S., **GÓTH, J.,** KIEZER, T., and SZLÁVIK, Z. (2004). An Entropy-Based Interpretation of Retrieval Status Value Based Retrieval, and Its Application to the Computation of Term and Query Discrimination Value. *JASIST (Journal of the American Society for Information Science and Technology)* Vol. 55, no 7, pp. 613-627 (SCI JIF = 2.086), ISSN: 1532-2882. [**thesis 2, 3**]

Paper in Hungarian journals:

[P4] DOMINICH S., **GÓTH, J.,** KIEZER, T., and SZLÁVIK, Z. (2003). NeuRadIR: Neuroradiológiai Információ-visszakereső Rendszer. *IME (Informatika és Menedzsment az Egészségügyben,)* II. évf. 1. szám, 2003. január-február, 41-45. oldal, ISSN:1588-6387.

Conference papers:

[P5] SZABÓ, T., **GÓTH, J.,** DOMINICH, S., KOZMANN, GY., SZOLGAY, P., and BÁRSONY, P. (2003). Novel Neuroradiological Image Processing and Information Retrieval in a Telestroke System. *Baud, R. at al.: The new navigators: from Professionals to Patients,* pp. 298-303, IOS Press, Amsterdam, ISBN: 1-58603-347-6.

[P6] **GÓTH, J.** (2002). Hyperbolic Information Retrieval. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, MF/IR,* pp. 61-77, Tampere, Finland, August 11-15.

[P7] DOMINICH, S., and **GÓTH, J.** (2002). Retrieval of Brain CT Reports and Images Using Interaction Information Retrieval. *Surjan, G. at al.: Health Data* *in the Information Society,* pp. 325-330, IOS Press, Amsterdam, ISBN: 1-58603-279-8.

6.3.2. Conference talks directly related to the thesis

International conferences:

1. SZABÓ, T., **GÓTH, J.,** DOMINICH, S., KOZMANN, GY., SZOLGAY, P., and BÁRSONY, P. (2003). Novel Neuroradiological Image Processing and Information Retrieval in a Telestroke System. *MIE2003 (XVIIth International Congress of the European Federation for Medical Informatics),* St. Malo, France, May 4-7.
2. **GÓTH, J.** (2002). Hyperbolic Information Retrieval. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, MF/IR,* Tampere, Finland, August 11-15.
3. DOMINICH, S., and **GÓTH, J.** (2002). Retrieval of Brain CT Reports and Images Using Interaction Information Retrieval. *MIE2002 (XVIIth International Congress of the European Federation for Medical Informatics),* Budapest, Hungary, August 25-29.

Hungarian conferences:

1. DOMINICH, S., **GÓTH,** J., KIEZER, T., and SZLÁVIK, Z. (2002). NEURADIR: Neuroradiológiai információ-visszakereső rendszer. *VEAB-NJSZT Orvosbiológiai Szakosztály közös rendezvénye*, Veszprém, 2002. december 11.
2. **GÓTH, J.,** KIEZER, T., and SZLÁVIK, Z. (2002). Képvisszakeresés az orvosi informatikában. *Matematikus, fizikus és informatikus doktorandusok 4. regionális találkozója.* Veszprémi Egyetem, Veszprém, 2002. május 30.

6.3.3 Other publications relevant to the thesis

[E1] DOMINICH, S., **GÓTH, J.,** HORVATH, M., KIEZER, T. (2005): Beauty of the World Wide Web—Cause, Goal, or Principle. in *David E. Losada, Juan M. Fernández-Lun., Advances in Information Retrieval, (Lecture Notes in Computer Science, Vol: 3408, Springer)* pp. 67-80 (SCI JIF = 0.515), ISBN: 3-540-25295-9.

[E2] DOMINICH, S., **GÓTH, J.,** and SKROP, A. (2003): A Study of the Usefulness of Institutions' Acronyms as Web Queries. in: *Sebastiani F., Advances in Information Retrieval, (Lecture Notes in Computer Science, Vol: 2633, Springer)* pp.580-587, (SCI JIF = 0.515), ISBN: 3-540-01274-5.

[E3] **GÓTH, J.** (2001). Szoftvertesztelés (Alapfogalmak, technikák). *Magyar Távközlés, XII. évfolyam, 2. szám,* 2001. február, 28-33 oldal. ISSN: 0865-9648.

6.3.4. Citations

1. Cited paper:

[P7] DOMINICH, S., and **GÓTH, J.** (2002). Retrieval of Brain CT Reports and Images Using Interaction Information Retrieval. *Surjan, G. at al.: Health Data in the Information Society,* pp. 325-330, IOS Press, Amsterdam, ISBN: 1-58603-279-8.

Citing paper:

BÁRSONY, P., and MAYER, I. (2002). Telestroke rendszer (agyérbetegségek intézetközi távkonzultációs rendszere). *IME (Informatika és Menedzsment az Egészésgügyben), I. évfolyam 5. szám* 2002. december, 28-32. oldal, ISSN: 1588-6387.

2. Cited paper:

[P5] SZABÓ, T., **GÓTH, J.,** DOMINICH, S., KOZMANN, GY., SZOLGAY, P., and BÁRSONY, P. (2003). Novel Neuroradiological Image Processing and Information Retrieval in a Telestroke System. *Baud, R. at al.: The new navigators: from Professionals to Patients,* pp. 298-303, IOS Press, Amsterdam, ISBN: 1-58603-347-6.

Citing paper:

ALBERTS, M., J., at al. (2005). Recommendation for Comprehensive Stroke Centers: A Consensus Statement From the Brain Attack Coalition. *Stroke* No 36, pp. 1597-1618, (SCI JIF = 5.748), ISSN: 0039-2499.

3. Cited paper:

[P4] DOMINICH S., **GÓTH, J.,** KIEZER, T., and SZLÁVIK, Z. (2003). NeuRadIR: Neuroradiológiai Információ-visszakereső Rendszer. *IME (Informatika és Menedzsment az Egészségügyben,)* II. évf. 1. szám, 2003. január-február, 41-45. oldal, ISSN:1588-6387.

Citing paper:

KOZMANN, Gy. (2005). Új információs technológiák az egészségügyben — Lehetőség a minőségi, gazdaságossági és versenyképességi elvárások teljesítésére. *IME (Informatika és Menedzsment az Egészségügyben,)* IV. évf. 3. szám, 34-39. oldal, ISSN:1588-6387.

4. Cited paper:

[P3] DOMINICH, S., **GÓTH, J.,** KIEZER, T., and SZLÁVIK, Z. (2004). An Entropy-Based Interpretation of Retrieval Status Value Based Retrieval, and Its Application to the Computation of Term and Query Discrimination Value. *JASIST (Journal of the American Society for Information Science and Technology)* Vol. 55, no 7, pp. 613-627 (SCI JIF = 2.086), ISSN: 1532-2882.

Citing paper:

IANEWA, T., Boldareva, L., Westerweld, T., Cornacchia, R., Hiemstra, D., and de Vries, A.P. (2004). Probabilistic approaches to video retrieval, *Proceedings of TRECVID Conference, National Institute of Standards*, NIST, USA.

5. Cited paper:

[P3] DOMINICH, S., **GÓTH, J.,** KIEZER, T., and SZLÁVIK, Z. (2004). An Entropy-Based Interpretation of Retrieval Status Value Based Retrieval, and Its Application to the Computation of Term and Query Discrimination Value. *JASIST (Journal of the American Society for Information Science and Technology)* Vol. 55, no 7, pp. 613-627 (SCI JIF = 2.086), ISSN: 1532-2882.

Citing paper:

LAFOUGE, T., Prime-Claverie, C. (2005). Production and use of information. Characterization of informetric distributions using effort function and density function. Exponential informetric process. *Information Processing and Management,* Vol. 41, pp. 1387-1394 (SCI JIF = 1.295), ISSN: 0306-4573.

Cumulated IF of publications: 4.301

Cumulated IF of citations: 7.043

REFERENCES

[1] Anderson, J.W. (1999). *Hyperbolic Geometry*. Springer Verlag, New York.

[2] Baclawski, K., and Simovici, D.A. (1996). A characterization of the information content of a classification. *Information Processing Letters*, vol. 57, pp. 211-214.

[3] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information retrieval*. ACM Press New York, Addison-Wesley.

[4] Belew, K.B. (2000). *Finding Out About.* Cambridge University Press.

[5] Berger, A., and Lafferty, J. (1999). Information retrieval as statistical translation. *Proceedings of the ACM SIGIR*, pp. 222-229.

[6] Berry, M.W. and Browne, M. (2000). *Understanding Search Engines – Mathematical Modeling and Text Retrieval.* SIAM, Philadelphia.

[7] Bolyai, J. (1987). *APPENDIX: The Theory of Space*. Akadémiai Kiadó, Budapest (Eds.: Kárteszi, F. and Szénássy, B.)

[8] Chu, H. and Rosenthal, M. (1996). Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology. *Proceedings of the American Society for Information Science Annual Meeting*, 33, pp: 127–135.

[9] Cooper, W.S., and Huizinga, P. (1982). The maximum entropy principle and its application to the design of probabilistic retrieval systems. *Information Technology, Research and Development*, 1, pp. 99-112.

[10] Crouch, C.J., and Yang, B. (1992). Experiments in automatic statistical thesaurus construction. *Proceedings of the 15th Annual International SIGIR Conference*, Denmark, pp:77-88.

[11] Császár, Á. (1974). *General Topology*. Akadémiai Kiadó, Budapest.

[12] Dominich, S. (2001). *Mathematical Foundations of Information retrieval*. Kluwer Academic Publishers, Dordrecht, Boston, London.

[12] Dominich, S. (2002). *Paradox-free Formal Foundation of Vector Space Model*. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information retrieval, Tampere, Finlad, August 11-15., pp. 43-60

[13] Dominich, S., Horváth, M., and Skrop, A. (2001). Evaluation of Interaction Information Retrieval. *Proceedings of the 23rd European Colloquium on*

Information retrieval Research, Springer Verlag, eWic, British Computer Society Information retrieval Specialist Group, GMD IPSI, Darmstadt, Germany, April 4-6, pp: 208-221

[14] Dubin, D. (1995). Document analysis for visualisation. *Proceedings of the Annual International ACM SIGIR Conference*, pp: 199-204.

[15] Fujii, A., and Ishikawa, T. (2001). Evaluating Multi-lingual Information retrieval and Clustering at ULIS. *Proceedings of NTCIR Meeting on Evaluation of Chinese & Japanese Text Retrieval and Summarization*, pp: 250-254.

[16] Gordon, M., and Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing and Management*, 35, pp: 141–180.

[17] Greiff, W.R., and Ponte, J.M. (2000). The Maximum Entropy Approach and Probabilistic IR Models. *ACM TOIS*, vol. 18, no. 3, pp. 246-287.

[18] Guazzo, M. (1977). Retrieval performance and information theory. *Information Processing and Management*, vol. 13, no. 3, pp. 155-165.

[19] Guy, C., and Fftyche, D. (2000). *An introduction to the principles of medical imaging*. Imperial College Press.

[20] Hilbert, D. and Cohn-Vossen, S. (1932). *Anschauliche Geometrie.* Springer Verlag, Berlin-Heidelberg-New York.

[21] Jansen, B. J., Spink, A., Bateman, J. and Saracevic, T. (1998a). Real life information retrieval: a study of user queries on the Web. *ACM SIGIR Forum,* **32**(1), pp: 5-17.

[22] Jansen, B. J., Spink, A. and Saracevic, T. (2000). Real life, real users and real needs: A study and analysis of users queries on the Web.*Information Processing and Management*, **36**(2), pp: 207-227.

[23] Kantor, P.B. (1984). Maximum entropy and the optimal design of automated information retrieval systems. *Information Technology*, vol. 3, no. 2, pp. 88-94.

[24] Kantor, P.B., and Lee, J.J. (1998). Testing the Maximum Entropy Principle for Information Retrieval. *Journal of the American Society for Information Science*, vol. 46, no. 6, pp. 557-566.

[25] Kolmogoroff, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Julius Springer, Berlin.

[26] Kowalski, G. (1997). Information retrieval systems: theory and implementation. Kluwer, London.

[27] Leighton, H. V., and Srivastava, J. (1999). First Twenty Precision among World Wide Web Search Services (Search Engines). *Journal of the American Society for Information Science,* **50**(10), pp: 870-881.

[28] Meetham, A. R. (1969). Communication theory and the evaluation of Information retrieval systems. *Information Storage and Retrieval*, vol. 5, pp: 129-134.

[29] Nigam, K., Lafferty, J., and McCallum, A. (1999). Using Maximum Entropy for Text Classification. *Proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering*, pp. 61-67.

[30] Oppenheim, C., Morris, A., and McKnight, C. (2000). The evaluation of WWW search engines. *Journal of Documentation*, **56**(2), pp: 190-211.

[31] Phillips, M. and Gunn, C. (1992). Visualizing hyperbolic space: Unusual uses of 4x4 matrices. In *1992 Symposium on Interactive 3D Graphics (Boston, MA, March 29 - April 1 1992)*, 25, pp: 209-214, New York. ACM SIGGRAPH. special issue of *Computer Graphics*.

[32] Phillips, M., Levy, S. and Munzner, T. (1993). Geomview: An interactive geometry viewer. *Notices of the American Mathematical Society*, **40**(8), pp:985-988, October 1993. Computers and Mathematics Column.

[33] Petterson, E.M. and Rutherford, D.E. (1965). *Einführung in die Abstracte Algebra*. Bibliographisches Institut, Mannheim.

[34] Ribeiro-Neto, B. and Muntz, R. (1996). A Belief Network Model for IR. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information retrieval, SIGIR'96*, pp. 253-260. August 18-22, 1996, Zurich, Switzerland

[35] Robertson, S.E, and Sparck Jones, K. (1977). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27, pp. 129-146.

[36] Salton, G. (1966). Automatic Phrase Matching. In Hayes, D.G. (ed.) *Readings in Automatic Language Processing*. American Elsevier Publishing Company, Inc., New York, pp. 169-188.

[37] Salton, G., Yang, C.S., and Yu, C.T. (1974). Contribution to the theory of indexing. *Information Processing 74*, North Holland Publishing Co., Amsterdam, pp: 584-590.

[38] Salton, G., Yang, C.S., and Yu, C.T. (1975). In theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, vol. 26, no. 1, pp: 33-44.

[39] Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw Hill, New York.

[40] Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM*, vol. 29, no. 7, pp: 648-656.

[41] Shannon, C. (1948). *A Mathematical Theory of Communication*. The Bell System Technical Journal, 27, pp: 379-423, 623-656, July, October.

[42] Sparck Jones, K. and van Rijsbergen, C.J. (1976). Progress in Documentation. *Journal of Documentation*, **32**(1), pp: 59-75.

[43] Spink, Amanda and Xu, Jack L. (2000). Selected results from a large study of Web searching: the Excite study. *Information Research*, **6**(1).

[44] Tan, C-M., Wang, Y-F., and Lee, C-D. (2002). The use of bigrams to enhance text categorization. *Information Processing and Management*, vol. 38, pp: 529-546

[45] Turtle, H. and Croft, W. B. (1991). Evaluation of an Inference Network-Based Retrieval Model. *ACM Transactions on Information Systems,* **9**(3), pp: 187-222.

[46] Turtle, H. and Croft, W. B. (1990). Inference Networks for Document Retrieval. *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp .1-24, Brussels, Belgium.

[47] Van Rijsbergen, C.J. (2004). *The Geometry of Information Retrieval*. Cambridge University Press, ISBN: 0521838053.

[48] Van Rijsbergen, C.J. (1979). *Information Retrieval.* Butterworth, London.

[49] Van Rijsbergen, C. J. (1987). *Információ visszakeresés.* Múzsák Közművelődési Kiadó, Budapest.

[50] Willett, P. (1985). An algorithm fro the calculation of exact term discrimination values. *Information Processing and Management*, vol. 21, no. 3, pp: 225-232.

[51] Xu, J. (1999) Internet search engines: real world IR issues and challenges. *Presentation to CIKM 99, October 31-November 4, 1999. Kansa City, MI.*

[52] Yoo, H.-W., Jang, D.-S., Jung, S-.H., Park, J.-H., and Song, K.-S. (2002). Visual information retrieval system via content-based approach. *Pattern Recognition*, **35**(3), pp. 749-769.

[53] Yu, C.T., and Salton, G. (1977). Effective Information Retrieval Using Term Accuracy. *Communications of the ACM*, vol. 20, no. 3, pp:135-142.

[54] ACTILYSE (Early CT Diagnosis of Acute Ischaemic Stroke: A Physicians' Guide), Boehringer Ingelheim, 1997.