Porto University

MAP-I PhD program

PhD Thesis Proposal

Improving the performance of text Information Retrieval (IR) Systems

Supervisor: Prof. Pavel Brazdil

By: Mohammadreza Vali Zadeh

Mar. 2012

Table of Contents

1. Introduction

Since the 1940s the problem of information storage and retrieval has attracted increasing attention. It is simply stated: we have vast amounts of information to which accurate and speedy access is becoming ever more difficult. One effect of this is that relevant information gets ignored since it has never uncovered, which in turn leads to much duplication of work and effort. With the advent of computers, a great deal of thought has been given to use them to provide rapid and intelligent retrieval systems. In libraries, many of which certainly have an information storage and retrieval problem, some of the more mundane tasks, such as cataloguing and general administration, have successfully been taken over by computers. However, the problem of effective retrieval remains largely unsolved [1].

The rapid and increasing development of the Internet, with the consequent huge availability of online textual information makes urgent the need for effective Information Retrieval Systems. The goal of an Information retrieval System (IRS) is to retrieve information considered pertinent to a user's query. The effectiveness of an IR System is measured through parameters, which reflect the ability of the system to accomplish such goals. However, the nature of the goal is not deterministic, since uncertainty and vagueness are present in many different parts of the retrieval process. The user's expression of his/her information needs in a query is uncertain, and often vague, the representation of a document informative content is uncertain, and so is the process by witch a query representation is matched to a document representation. The effectiveness of an IR System is therefore crucially related to the system's capability to deal with the vagueness and uncertainty of the retrieval process. Commercially available IR Systems generally ignore these aspects.

In recent years a great deal of research in IR has aimed at modeling the vagueness and uncertainty, which invariably characterize the management of information. A First class of approaches is based on methods of analysis of natural language [2]. The main limitation of these methods is the level of deepness of analysis of the language, and their consequent range of applicability; a satisfying interpretation of documents meaning needs a too large number of decision rules even in narrow application domains. A second class of approaches is Probabilistic IR. However, there is another

set of approaches receiving increasing interest. This set of approaches goes under the name of Soft Information Retrieval or intelligent IR.

2. Information Retrieval

Information Retrieval is a Branch of Computing Science that aims at storing and allowing fast access to a large amount of information. This information can be of any kind: textual, visual, or auditory [3].

An information retrieval system is a system that is capable of storage, retrieval, and maintenance of information items. Presently, most retrieval of non-text items is based on searching their textual descriptions. Text items are often referred to as *documents,* and may be of different scope (book, article, paragraph, etc.). Most actual IR Systems store and enable the retrieval of only textural information or documents. However, this is not an easy task, just to give a clue to its size, it must be noticed that often the collections of documents an IR System has to deal with contain several thousands or sometime millions of documents.

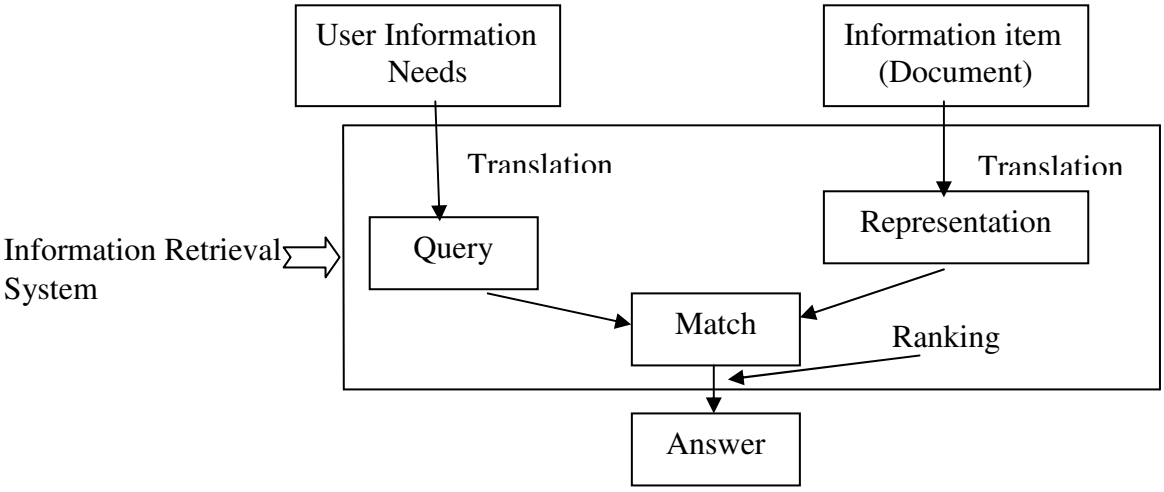

Fig. 1 Schematic overview of an Information Retrieval System

A user accesses the IR System by submitting a query; the IR System then tries to retrieve all documents that are relevant to the query [15]. To this purpose, in a preliminary phase, the documents contained in the archive are analyzed to provide a formal representation of their contents: this process is known as “indexing”. Once a document has been analyzed, a surrogate describing the document is stored in an index, while the document itself is also stored in the collection or archive. To express some information needs a user formulates a query, in the system’s query language.

The query is matched against entries in the index in order to determine which documents are relevant to the user.

In response to a query, an IRS can provide either an exact answer or a ranking list of documents that appear likely to contain information relevant to the query. The result depends on the formal model adopted by the system. As it will be explained in the following sections, the Boolean model produces an extract answer, while others, apply a partial matching mechanism, which produces a ranking of the retrieved documents so that the documents most likely to be relevant are presented to the user first. In some IRSs queries are expressed in natural language and to be processed by the system they are passed through a query processor which breaks them into their constituent's words. Stop words and suffixes are removed, so that what remains to represent query and documents are lists of terms that can be compared using some "relevance evaluation" algorithms. A scheme of an IR system is depicted in Fig. 1.

2-2 Information Retrieval Models

The choice of the formal background to define both the document and query representations characterizes the model of an IRS. In the IR literature different models have been proposed.

The Boolean model is still the one most commonly used in commercial IR systems. It is based on mathematical set theory. Here documents are represented as sets of index terms and Boolean search strategy retrieves those documents which are 'true' for the query. This formulation only makes sense if the queries are expressed in terms of index terms (or keywords) and combined by the usual logical connectives AND, OR, and NOT. For example, if the query Q = ($K1$ AND $K2$) OR ($K3$ AND (NOT $K4$)) then the Boolean search will retrieve all documents indexed by $K1$ and $K2$, as well as all documents indexed by $K3$ which are *not* indexed by $K4$ [1].

The vector space (statistical) model is based on a spatial interpretation of both documents and queries. An improvement of the document representation over the Boolean model is obtained by associating with each index term a numeric value, called the index term weight, which expresses the variable degree of significance that the term has in synthesizing the information content of the document. Similarity measures (matching function) between document and query representation are then used to evaluate a document's relevance with regards to a query.

There are many examples of matching functions in the literature. Perhaps the simplest is the one associated with the simple matching search strategy.

If *M* is the matching function and D is the set of keywords representing the document, and *Q* the set representing the query, then:

$$M=\frac{2 \| D \cap Q \|}{\| D \| + \| Q \|}$$

A popular one used by the SMART project, which they call cosine correlation, assumes that the document and query are represented as numerical vectors in *t*-space, that is $Q = (q1,q2, \ldots, qt)$ and $D = (d1, d2, \ldots, dt)$ where *qi* and *di* are numerical weights associated with the keyword *i*. The cosine correlation is now simply:

$$r = \frac{\sum_{i=1}^{t} q_i d_i}{\left(\sum_{i=1}^{t} (q_i)^2 \sum_{i=1}^{t} (d_i)^2 \right)^{1/2}}$$

Or, in the notation for a vector space with a Euclidean norm,

$$r = \frac{(Q, D)}{\| Q \| \times \| D \|} = \text{cosine}\, \theta$$

Where θ is the angle between vectors *Q* and *D*.

The Probabilistic Model [3] ranks documents in decreasing order of their evaluated probability of relevance to a user's information need. Research has made much use of formal theories of probability and of statistics in order to evaluate, or at least estimate, the probability of relevance. The task of a probabilistic IR system is to rank documents according to their estimated probability of being relevant.

2-3- Relevance Feedback and Query Expansion

The word feedback is normally used to describe the mechanism by which a system can improve its performance on a task by taking into account the past performance. In other words a simple input-output system feeds back the information from the output so that this may be used to improve the performance on the next input. The notion of feedback is well established in biological and automatic control systems. It has been popularized by Norbert Wiener in his book *Cybernetics* [1]. In information retrieval Relevance feedback has been used with considerable effect [3, 4,14].

Relevance feedback is a technique that allows a user to interactively express his information requirement by modifying his original query formation with further information. This additional information is often provided by indicating some relevant documents among the documents retrieved by the system. When a document is marked as relevant the RF device analyses the text of the document, picking out terms that are statistically significant, and adds these terms to the query.

Obviously the user cannot mark documents as relevant until some are retrieved, so the first search has to be initiated by a query and the initial query specification has to be good enough to pick out some relevant documents from the collection. If at least one document in the list of retrieved documents matches, or come close to match, the user can mark the document(s) as relevant and starts the RF process. If RF performs well the next list should be closer to the user's requirement and contain more relevant documents [4]. A schematic view of a RF device is depicted in Fig. 2.

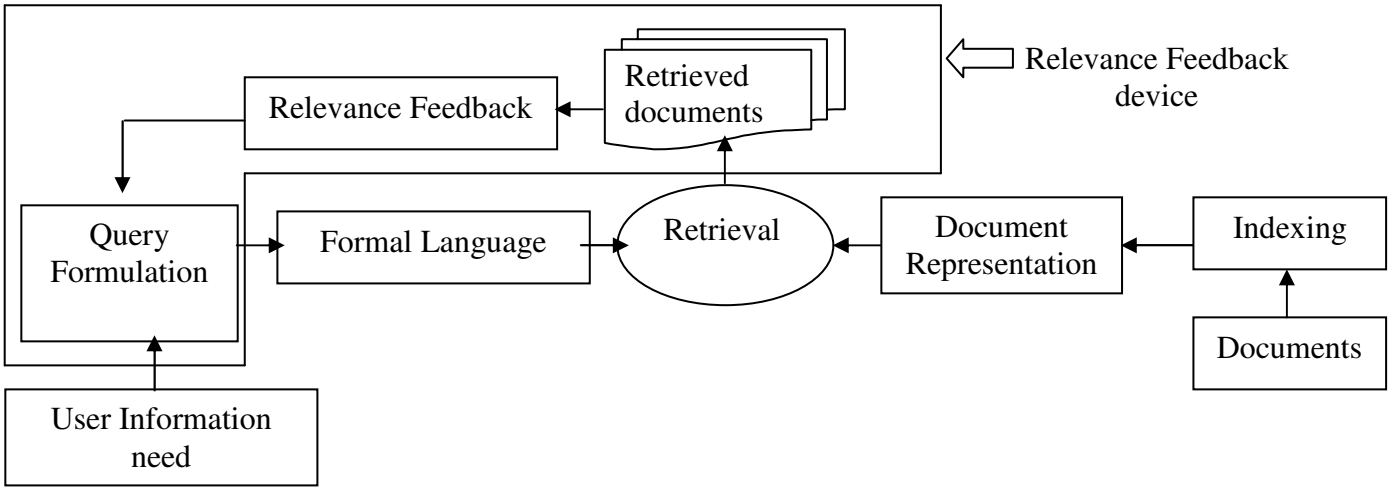

Fig.2. Schematic view of a relevance feedback device.

2-3- Clustering based IR systems

"Clustering" of documents is the grouping of documents into distinct classes according to their intrinsic (usually statistical) properties to improve retrieval efficiency. Clustering also improves retrieval effectiveness based on the *cluster hypothesis:* closely associated documents tend to be relevant to the same queries [5, 6].

Clustering is a kind of classification but it differs from the classification for routing purposes. In a routing application, the documents are classified in terms of their similarity or relevance to external queries or topics or user profiles. In

"clustering," we seek features that will separate the documents into natural groups based entirely on the internal properties of the collection. Ideally, the groups will be completely separate and as far apart as possible in feature space. But sometimes, overlap of clusters is unavoidable. [3] Since clustering depends on the statistical properties of the collection being clustered rather than on matching the documents against some external set of queries, it is normally (but not always) applied to a pre-existing collection rather than an incoming stream of documents as in a routing application [3, 7].

A cluster-based search proceeds to satisfy a query efficiently by identify and retrieving only those clusters which exhibit a sufficiently high degree of match with the query. Clustering improves the effectiveness of retrieval as it results in the retrieval of a higher number of relevant documents for a given amount of effort [3].

2-3-1 Clustering methods

In Information Retrieval Systems many methods of clustering are proposed. These methods are explained briefly in the following sections.

2-3-1-1 Hierarchical Clustering

A *hierarchical* algorithm may start by considering all the documents as a single cluster and then breaking it down into smaller clusters ("divisive" clustering). The algorithm can start with the individual documents and group them together into progressively larger clusters "agglomerative" clustering. Agglomerative clustering produces a hierarchy of clusters grouped into larger clusters. It is often called Agglomerative Hierarchical Clustering or AHC for short. In the latter case, the similarities are sorted in descending order. Initially, each document is considered a separate cluster. The general rule is that at each stage the two most similar clusters are combined. Initially, the most similar documents are combined into a cluster. At that stage, "most similar" means having the highest similarity of any document pair. Thereafter, we need a criterion for deciding what "most similar" means when some of the clusters are still single documents and some are multi-document clusters that we have previously formed by agglomeration (or when all of the clusters have become multi-document). The various agglomerative cluster methods are distinguished by the rule for determining inter-cluster similarity when one or both clusters contain multiple documents [7, 8, 9].

In “single-link” clustering (the most famous clustering method), the similarity between two clusters is defined to be “the similarity between the *most similar* pair of items, one of which appears in each cluster; thus each cluster member will be more similar to at least one member in that same cluster than to any member of another cluster.” The algorithm is called “single-link” because two clusters can be combined on the basis of one high similarity between a document in the one cluster and a document in the other. It is also called “nearest neighbor” clustering because two clusters are combined on the basis of the two documents, one from each cluster, that are nearest to each other. Hence, each cluster is formed by a chain of nearest neighbor document-to-document single links.

In “complete-link” clustering by contrast, the similarity between two clusters is defined to be “the similarity between the least similar pair of items” one of which appears in each cluster. Thus each cluster member is more similar to the most dissimilar member of that cluster than to the most dissimilar member of any other cluster”. Hence, in the “complete-link” algorithm, the similarity between two clusters (which determines whether they should be combined or not) depends on *all* the similarities between documents in the one cluster and documents in the other.

2-3-1-2 Heuristic Clustering

The term “heuristic” has been used by authors such as Rijsbergen [3] to characterize methods that take shortcuts to achieve greater efficiency in terms of space and time requirements. In particular, such terms refer to cluster methods that do *not* generate or do not access the full *O (N^2)* set of inter-document similarities in a collection of *N* documents. Such methods effectively make fewer (sometimes far fewer) effective “passes” through the inter-document similarity matrix or its equivalent. In particular, it is characteristic of many such methods that the clusters for a given set of *N* documents vary depending on the order in which documents are initially referenced.

2-3-1-3 Incremental Clustering

Incremental methods make use of a similarity measure but they don’t require that similarities be pre-computed for all document pairs. Indeed, all document pairs are not available initially, since by definition, incremental methods cluster a stream of

incoming documents. The similarities are computed “on the fly” as the documents stream past the incremental cluster system [10, 11].

3- Intelligent Information Retrieval

In recent years a lot of effort has been devoted to improve the performance of IR systems and research has explored many different approaches to use Machine Learning techniques in improving IR systems.

Machine learning has many branches: neural network, evolutionary algorithms, decision tree and etc. Genetic algorithm is a Machine Learning and optimization tool which can be used for improving the performance of information systems.

Evolutionary algorithms are based on the Darwinian principles of natural selection. These algorithms can be further divided into: GA's, evolutionary strategies, and evolutionary programming. While evolutionary programming utilizes changes at the level of species, the evolutionary strategies, exploit changes at individual behavioral level.

GA's are based on genetic operators of selection, crossover, and mutation. GA's are robust in searching a multidimensional space to find optimal or near optimal solutions. There are a few studies in IR literature that incorporate GA to improve the performance of IR systems [8, 12, 13].

4- Thesis Goal

As mentioned above, there is a growing need for research at the intersection of Information Retrieval, Clustering and Artificial Intelligence. The most important problem in IR Systems is that a single word may have many meanings and different people may use different words for a single meaning. Use of thesaurus and query expansion techniques aims at improving this problem.

The aim of this project is to improve the performance of IR systems based on machine learning techniques and clustering. Special attention will be made to fine-tune an IR system based on clustering, adaptation and learning. The work will be in the following:

Upgrading the performance of clustered-based IR systems with the aid of soft computing techniques.

I try to use a new query sensitive similarity measure for reranking in IR system to improve its performance.

Proposing a model for combining ranked IR systems and clustering based IR systems.

Also, I try to tune feedback method by query sensitive similarity measures. We know that feedback in IR systems can improve its performance, and there are some methods that are proposed and used by other researchers. I try to propose a new method and evaluate it.

The final try, using the query sensitive similarity measure in multi text summarization.

The effectiveness of proposed methods will be evaluated using precision and recall on standard data collections.

5- Work plan

4 months	Primary study on the subject of Information Retrieval and Clustering.
2 months	Write the proposal and defence.
4 month	Study the work done by other researchers in the field of adaptive Information retrieval Systems.
4 months	Design and propose new methods for reranking and clustering.
4 months	Implementation
3 months	Evaluation of the performance of the method proposed using standard methods.
4 months	writing the first thesis and pre defence.
3 months	Correcting the thesis and defence.

6- References

1- van Rijsbergen, C. J. ,*on line book*, Http://www.dcs.gla.ac.uk/Keith/Preface.html

2- Smeaton, A.F. *Progress in the application of natural Language Processing to Information Retrieval tasks*, The Computer Journal, 35, 1992, 268-278.

3- van Rijsbergen, C. J., *Information Retrieval*, Butterworths, London, second edition, 1979.

4- Buell,D.A. and Kraft, D.H. *A model for a weighted retrieval system, Journal of the American Society for Information Science*, 32(43), May, 211-216, 1981.

5- Anastasios Tombros, Robert Villa, C.J. van Rijsbergen, *The Effectiveness of Query-Specific Hierarchic Clustering in Information Retrieval*, Information Processing and Management 38(4): 559-582 (2002).

6- Anastasios Tombros, C.J. van Rijsbergen, *Query-Sensitive similarity measures for the Calculation of Interdocument relationships*, CIKM 2001, 17-24.

7- Ed Greengrass, *Information Retrieval: A Survey*, November 2000, http://www.csee.umbc.edu/cadip/readings/IR.report.120600.book.pdf

8- Pathak, Praveen and Fan, weiguo and Gordon, Michael D., *personalization of search engine services for effective retrieval and knowledge management,* Http://aisel.isworld.org/password.asp?Vpath=ICIS/2000&PDFpath=00crp04.pdf

9- Allan, J., J. Callan, W. B. Croft, L. Ballesteros, D. Byrd, R. Swan, & J. Xu (1998). Inquery does battle with TREC-6. In *Sixth Text Retrieval Conference (TREC-6)*, pp. 169–206.

10- J. Koenemann and N. J. Belkin. A case for interaction: *A study of interactive information retrieval behavior and effectiveness*. In Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems, pages 205-212, 1996.

11- Hearst, M.A. and Pedersen, J.O. (1996). *Re-examining the cluster hypothesis: Scatter/Gather on Retrieval Results. In Proceeding of the 19th Annual ACM SIGIR Conference, pp.* 76-84. Zurich, Switzerland.

12- Pathak P., Gordon M., Fan W., (1999), *Effective information retrieval using genetic algorithms based matching functions adaptation.* Journal of the American Society for Information Science, v.50 n.9, p.760-771, July 1999.

13- Goldberg D.E., *Genetic algorithms in search, optimization and Machine Learning* Addison-Wesley, 1989.

14- Guido Zuccon, Leif Azzopardi, C. J. van Rijsbergen: An Analysis of Ranking Principles and Retrieval Strategies. ICTIR 2011: 151-163,2011

15- Weiguo Fan, Praveen Pathak, Mi Zhou: Genetic-based approaches in ranking function discovery and optimization in information retrieval - A framework. Decision Support Systems 47(4): 398-407 (2009)

Journals

- Information Retrieval, Subject: management of computing and information system, data structures, publisher: Springer Netherlands
- Transactions on Information Systems, ACM Transactions on information systems
- Information Processing Letters, Elsevier, Impact Factor:0.774
- Information Sciences, Elsevier
- Information Processing & Management
- American Society for Information Science and Technology

Conferences

-Research and Development in information retrieval

-International conference on information and knowledge management

- European colloquium on IR research

- Text REtrieval conference (TREC)

- ACM and IEEE Joint conference on digital libraries

- International conference on Asian digital libraries

- Information interaction in context

Mohammadreza Vali Zadeh

Pavel Brazdil

(Pavel Brazdil)