SETIT 2005
3rd International Conference: Sciences of Electronic,
Technologies of Information and Telecommunications
MARCH 27-31, 2005 – TUNISIA

# Intelligent Information Retrieval

Pooja Jain
Indian Institute of Information Technology
Allahabad

**Abstract**

In this paper an intelligent agent-based model for information retrieval is presented. The growing amount of on-line information and its dynamic nature forces us to reconsider existing passive approaches for information retrieval. Because of this ever-growing size of information sources the burden of retrieving information cannot be simply left on users. Our approach uses agent-based paradigm in order to handle this problem. Further in order to avoid users being overloaded with bulk of irrelevant information along with relevant ones and to improve ranking of the returned documents, we attempt to include semantics in making relevance judgment through conceptual graphs. We have first applied vector space model and then used conceptual graph to obtain final ranking. The results achieved show improved ranking of the returned documents.

## 1. Introduction

Most of he search engines being used currently do return a lot of irrelevant information that do not meet user's requirements. Users are not interested in huge amount of information, but in precise, accurate and relevant information. The success of existing search engines depends on the appropriate choice of keywords by user in framing their queries. Most of the search engines use keyword matching. With keyword matching approach it is not possible to distinguish among relevant and irrelevant document if documents use similar terms but in different context. Another drawback of existing search and retrieval engines is their passive mode of working. They retrieve information only when they are asked to do so. With ever-growing size of information sources this passive mode of working is undesirable, as the amount of on-line information being made available every day is beyond the capacity of a single user. This forces us to consider alternate model for information retrieval. This alternate model should be autonomous and proactive [1], so as to free the users to focus their attentions to other important tasks. This should include context information for making relevance judgment, so as to avoid users from bulk of irrelevant documents and should be intelligent.

In this paper a novel approach for information retrieval has been presented that combines the use of conceptual graphs and multi-agent paradigm. It consists of user modeling agents, retrieval agents and a facilitator.

## 2. Overview of Background Topics

### 2.1. Intelligent agent

Intelligent agent technology has its roots in the idea that patterns of behavior can be identified and described. Computers are able to follow rules. If we can state our rules and patterns to computer then we can design systems that can carry out actions based on these rules.

However it is difficult to explain how to recognize patterns automatically. In a specific domain, to some extent it is possible to teach computers to find patterns, extract rules and implement them. Computer programs can be written which can trace user's activities and can identify actions and group them into class of action. That is precisely what an intelligent agent is supposed to be i.e. Computer programs, which can learn patterns of behavior and then act on behalf of user. In context of information retrieval this means that such an agent can satisfy the user's quest for information by identifying his interests and preferences and free him to some extent so that he can concentrate on the real intellectual task of reading the documents. Intelligent agents are being promoted as the next generation model for engineering complex, distributed systems. There are many definitions of an "intelligent agent" but most current researchers do agree on the following criterion –

### 2.2 Vector space model

In vector space model both document and query are represented as vectors. Retrieval is performed on the basis of "closeness" of 'query vector' and 'document vector'. Given a set of n documents –

$$D = \{d_1, d_2, d_{3, \ldots,} d_{j, \ldots} d_n\}$$

And a finite set of terms

$$T = \{t_1, t_2, t_3, \ldots, t_i, \ldots t_m\}$$

any document dj will be represented by a vector $v_j$ as follows –

$$v_j = (w_{1j}, w_{2j}, w_{3j}, \ldots, w_{ij}, \ldots, w_{mj})$$

where $0 <= w_{ij} <= 1$, is weight of term $t_i$ in document $d_j$.

The tf-idf weighting scheme is used. With this scheme weight of term $t_i$ in document $d_j$ defined as the product of term frequency (tf) and inverse document frequency(idf). Inverse document frequency favors terms that occur in fewer documents. It is calculated as –

$$idf_i = \log \frac{n}{n_i}$$

where $n_i$ = no. of documents in which term $t_i$ occurs. Thus weight of any term $t_i$ will be given by –

$$w_{ij} = tf_{ij} \times idf_i$$

The similarity measure used is well-known dice's coefficient [8], given as follows –

$$S_{j,k} = \frac{2(v_j, v_k)}{\sum_{i=1}^{m} (w_{ij}, w_{ik})}$$

where , $v_j$ = document vector and $v_k$ = query vector

There are two accepted standards for comparing and evaluating performance of information retrieval systems, precision and recall, defined as

$$\text{Precision} = \frac{\text{No.of relevant Documents retrieved}}{\text{Total no. of retrieved documents}}$$

## 2.3 Latent Semantic indexing

$$\text{Re}\,call = \frac{\text{No. of relevant Documents retrieved}}{\text{Total no. of relevant documents}}$$

The Latent Semantic Indexing information retrieval model builds upon the prior research in information retrieval and uses the singular value decomposition (SVD) to reduce the dimensions of the term-document space. LSI explicitly represents terms and documents in a rich, high-dimensional space, allowing the underlying (``latent''), semantic relationships between terms and documents to be exploited during searching.

LSI differs from other methods at using reduced-space models for information retrieval in several ways. Most notably, LSI represents documents in a high-dimensional space. Secondly, both terms and documents are explicitly represented in the same space. Thirdly no attempt is made to interpret the meaning of each dimension. Each dimension is merely assumed to represent one or more semantic relationships in the term-document space. Finally, LSI is able to represent and manipulate large data sets, making it viable for real-world applications

## 2.4. Conceptual Graphs

Conceptual graphs are highly expressive form of logic and were originally designed for representing natural language semantics. They have been evolved out of conceptual structure theory as set down by Sowa [4]. Sowa defines Conceptual graph as follows -

"Conceptual graphs form a knowledge representation language based on linguistics, psychology and philosophy. In this graph concept node represent entities, attributes, states and events and relation nodes show how the concepts are interconnected."

Concept nodes consist of a type field and a referent field. A blank referent field implies presence of existential quantifier by default. Details can be found in [6].

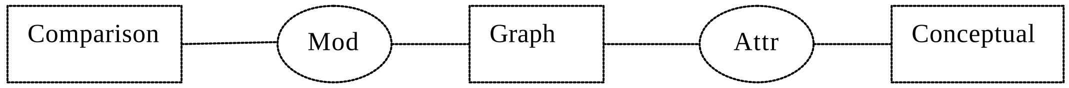
Conceptual graphs are very closely related to natural language and hence can be used for representing text. Such a representation holds the promise of extracting more information from documents by explicitly capturing logical relationship between objects, unlike word-statistical approaches that merely count nouns and noun phrase. This fact suggests use of conceptual graphs in information retrieval. With such representation we will be able to improve precision in information retrieval.

Sesei CG builder* has been used for constructing conceptual graphs of pieces of text. The output of this tool is in CGIF (conceptual graph interchange format). For example, if input is "comparison of conceptual graph" the output will be –*

[Comparison *a] [Graph *b] [Conceptual *c] (mod?a ?b) (attr ? b ?c)

The output is CGIF representation of the following conceptual graph.

* CG-builder component of **sesei** ( SEmantic Search Engine for the Internet) , available online



Linear notation of the above CG is -

[comparison] → (mod) → [graph]→ (attr) → [conceptual]

We have used similarity measure given by Montez [9] for comparing Conceptual graph representation of text. Given two texts represented by the conceptual graphs *G*1 and *G*2 respectively and one of their intersection graphs $G_c$, similarity *s* between them will be a combination of their conceptual similarity $s_c$ and relational similarity $s_r$ given as –

$$s = s_c \times (a + b \times s_r)$$

$$s_c = \frac{2 \times n(G_c)}{n(G_1) + n(G_2)}$$

$$s_r = \frac{2 \times m(G_c)}{mG_c(G_1) + mG_c(G_2)}$$

$$a = \frac{2 \times n(G_c)}{2 \times n(G_c) + mG_c(G_1) + mG_c(G_2)}$$

and b = 1 - a

where $m(G_c)$ is the number of the arcs in the graph $G_c$, $mG_c(G)$ is the number of the arcs in the immediate neighborhood of the graph $G_c$ in the graph G.

## 3. Application discussion

To understand the importance of the intelligent multi agent information systems, we will consider a system that is used to retrieve information from document database. Our database consists of papers from various disciplines. User will present query through an interface, user modeling agent will modify the query based on user's profile. Modified query is then passed to facilitator which has knowledge about database and routes query to appropriate retrieval agent. The retrieval agent will first retrieve a set of documents using vector model and then apply conceptual graph to make relevance judgment and finally documents are returned to user.

## 4. System Structure

### 4.1 Architecture of the System

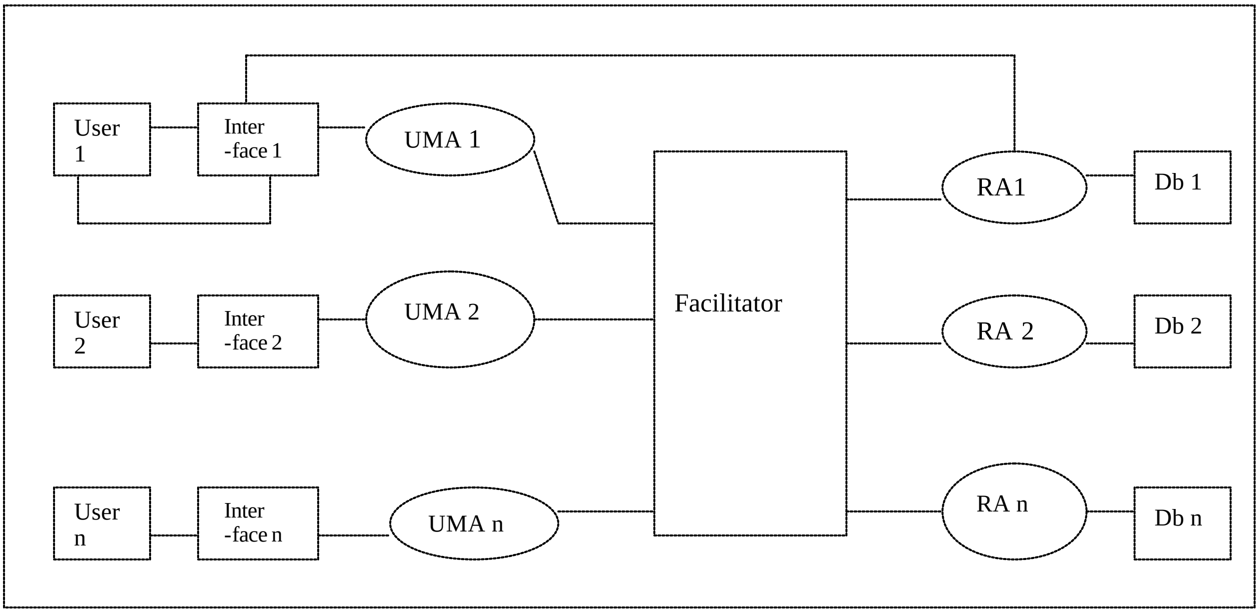


**Fig 1 .**A-User Modeling Agent, RA-Retrieval Agent, Db-Database

The system consists of user modeling agents, a facilitator module, retrieval agents and an interface module. The **user modeling agent** is used to model specific user to which it is connected. A one to one mapping is assumed for user modeling agent. It has the learning capability and stores the details of the user. It is responsible for interacting with the user regarding relevance of the document. The facilitator **acts** as mediator between user and information resources. A facilitator is associated with a group of agents. Association between facilitator and agent groups occurs when application starts and when a user exits from application its information is removed from facilitator's local database. Facilitator forms a virtual path between the user and the agent that can provide the information to the user. When such a path has been created then the user can directly query the agent in a more detailed manner and can get the desired data from the database. The **retrieval agent** performs the necessary matching and then gives the results of the query to the user after consulting the database. If a new database is added, it gives its information to the facilitator, so that it can analyze the query properly.

### 4.2 User modeling agent(UMA)

User modeling agent is responsible for construction of user's profile. A user profile concerns the stored knowledge maintained by the system about the user's interest [10]. Initially system does not have much information about users. Their profile will build up gradually. The information that will be collected about user at the time of registration includes username, password, area of interest, preferred sites etc. Following steps explain the functions of UMA –

1. User makes a query
2. User modeling agent (UMA) running in the background will notice this activity.
3. A log of this query is maintained by UMA.
4. When results corresponding to the above query are returned, feedback on the relevance is obtained from user.
5. User modeling agent extracts semantics of the most relevant document, and updates profile accordingly.
6. UMA may interact directly with the user to get more information.

### 4.3 Retrieval agent (RA)

Retrieval agent is responsible for retrieving information from the database. The basic functionality of this agent can be described as –

1. Accepts queries from facilitator component of interface agent.
2. Match the query against documents in the database.
3. Retrieve documents and returns it to query interface or user.

### 4.4 Facilitator

It is the interface between user and rest of the system. Most of the interaction of the user with the system is through this interface. It's working can be better explained with following steps –

1. Accepts query
2. Routes query to appropriate Database
3. Collect information about database.

## 5. Conclusions

In this way we can say that the intelligent multi agent system is able to give the information in a much refined way and works actively. As the approach presented in this paper deals with semantics we are able to get improved precision. Our approach is thus capable of facing the challenges posed by explosive growth of information and on-line databases.

## 6. Acknowledgements:-

Dr.U.S.Tiwary, Indian Institute of Information Technology
Tanveer J. Siddiqui, J.K. Institute of Applied Physics & Technology,University of Allahabad, Allahabad

## 7. References

[1]. Wondergem, B. C. M., Bommel, P. Van, Huibers, T. W. C. and Weide, Th. van der " Towards an agent based retrieval engine ( Profile – Information Filtering Project)", In furner and Harper editors, Proceedings of the 19th BCS-IRSG Annual Colloquium on IR Research, pages 126-144, Aberdeen, Scotland, April 1997, Robert Gordon University.

[2]. Wooldridge, M. and Jennings, N. R., (1995) " Intelligent agents: Theory and Practice, The knowledge engineering review", 10(2): 115-152

[3]. Susan Feldman, Dataserach and Edmund Yu, "Intelligent agents: A Primer", Searcher, Vol. 7, No. 9, October 1999.

[4]. Sowa, J. F. (1984) "Conceptual structures – Information processing in mind and machine", Addison –Wesley.

[5]. Salton, G. McGill M. J. (1983) Introduction to modern Information retrieval, McGraw-Hill.

[6]. Sowa, J. F.(1993) "Relating diagram to logic", In Eds. Guy W. Mineau, Bernard Moulin and John F. Sowa, Proceedings of first International conference on conceptual structures, ICCS'93, Quebec city, Canada, August 4-7.

[7]. Hongchen Fu (2001), "Intelligent Agents and its Applications in Information Retrieval".

http://mcs.open.ac.uk/hf35/computing/MSC/AI/ai.pdf.

[8]. Rasmussen, Edie (1992). "Clustering Algorithms". Information Retrieval: Data Structures & Algorithms. William B. Frakes and Ricardo Baeza-Yates (Eds.), Prentice Hall, 1992.

[9]. Montes-y-Gómez, M., A. López-López, A. Gelbukh (2000) "Comparison of conceptual graphs", O. cario, L. E. Sucar, F. J. Cantu(Es.) MICAI 2000: Advances in Artificial Intelligence. Lecture notes in Spriger –Verlag, pp. 548-556, 2000.

[10]. María J. Martín-Bautista , Henrik L. Larsen , Daniel Sánchez, María-Amparo Vila, "Information Filtering and User Profile Construction with Fuzzy Sets and Genetic Algorithms", Paper downloaded from Internet

[11]. R. Patil, R. Fikes, P. Patel-Schneider, D. McKay, T. Finin, T. Gruber and R. Neches. "The DARPA Knowledge Sharing Effort : Progress Report" In Principles of Knowledge Representation and Reasoning : Proceedings of the Third International Conference, Nov. 1992, Available as http://www.cs.ubmc.edu/kqml/-papers.ps

[12].Tim Finin, Richard Fritzson, Don McKay and Robin McEntire, "KQML as an Agent Communication Language", In The Proceedings of the Third International Conference on Information and Knowledge Management (CIKM'94), ACM Press, November 1994.