International Journal of communication and computer Technologies, ISSN: 2278-9723

Available at http://www.ijccts.org

Search Accuracy in Web Information Retrieval

Trilok Gupta[1], Archana Sharma[2]

[1]*Pacific Academy of Higher Education and Research University, Udaipur, Rajasthan-India,*

[2]*Associate Professor, Department of Computer Science, Gurukul Institute of Engineering & Techchnology Institutional Area,*

Received: 16-06-2014, **Revised:** 02-09-2014, **Accepted:** 11-10-12014, **Published online:** 01-12-2014

Abstract: Due to the exponential growth of the Internet in recent years, search engines have the complex task of sorting through billions of pages and displaying only the most relevant pages for the submitted search query. The aim of this paper is to address the above problems in order to improve retrieval accuracy in Web information retrieval. By indexing a target Web page more accurately, and allowing each user to perform more fine-grained search this satisfy his/her information need.

Keywords – Word Wide Web, Information Retrieval, Indexing, Paid listings, Page Rank, pay per click, click-through-rate,

I. INTRODUCTION

Modern search engines are pretty incredible complex algorithms enable search engines to take search query and return results that are usually quite accurate, presenting with valuable information, suggest amidst a vast information data mine.

The history of the Internet begins during the 1950s and 1960s with the development of computers. It came about as a result of early visionaries who saw a great value sharing scientific and military research information via computers.

With the launch of the Sputnik by the USSR in 1957, the United States established the Advanced Research Projects Agency (ARPA) with the goal of becoming a leader in science and developing new technologies. In 1962, Dr. J.C.R. Licklider was chosen to lead ARPA's research efforts and was a key figure in laying the foundation for ARPANET, which would eventually become the Internet.

It wasn't until December of 1969 that it was brought online, and at the time, there were only four computers connected at the following universities: UCLA, Stanford, UCSB and the University of Utah; you can see the original four-node network in Figure 1.1.

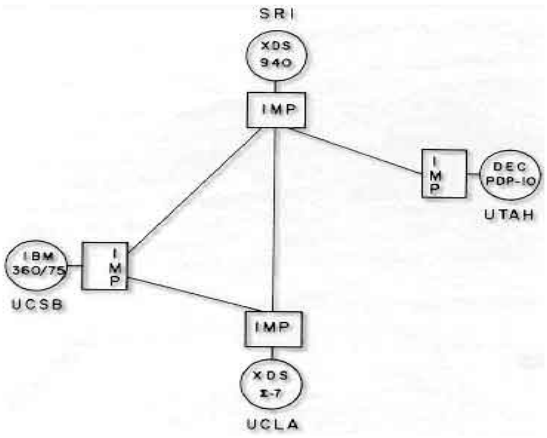

Figure 1.1: ARPANET: Four-node network in 1969.

In 1990, Tim Berners-Lee created the first Web browser (and Web editor) originally called the World Wide Web and later renamed to Nexus in order to avoid "confusion between the program and the abstract information space (which is now spelled World Wide Web with spaces)" [1]; it was written in Objective-C using the NeXT computer. And at the time, this was the only way to browse the web. You can see a screenshot of the first browser in Figure 1.2 below.

1993 marked an important turning point

for the World Wide Web. The National Centre for Supercomputing Applications (NCSA) at the University of Illinois, led by Marc Andreessen, introduced the Mosaic browser. It quickly became popular due to its graphical support and its ability to "display images inline with text instead of displaying images in a separate window" [1]. Mosaic made it much easier for people to navigate hyperlinked pages and it made the Web "easy to use and more accessible to the average person. Andreesen's browser sparked the internet boom of the 1990s" [3].

business and released Internet Explorer which was "heavily influenced by Mosaic, initiating the industry's first browser war Bundled with Windows, Internet Explorer gained dominance in the web browser market" [4].

Alta Vista was the first search engine to process natural language queries; Lycos started strong with a system categorizing relevance signals, matching keywords with prefixes and word proximity; and Ask Jeeves introduced the use of human editors to match actual user search queries.

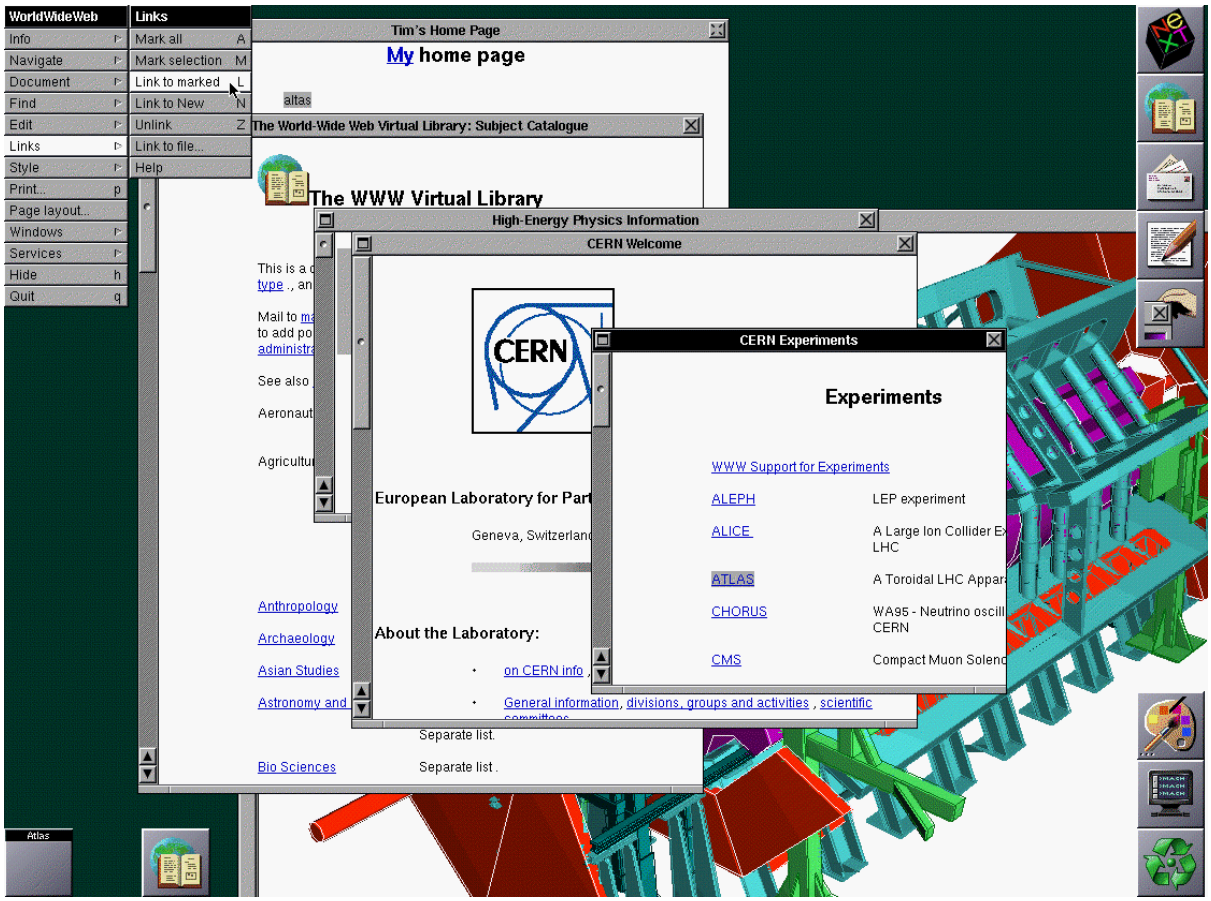

Figure 1.2: Screenshot of the first web browser called World Wide Web launched in 1990.

A year later, Andreesen "started his own company, named Netscape, and released the Mosaic-influenced Netscape Navigator in 1994, which quickly became the world's most popular browser, accounting for 90% of all web use at its peak" [4]. Then in 1995, Microsoft got involved in the web browser

Searching online has become part of the everyday lives of most people, whether to look for information about the latest gadget to getting directions to a popular restaurant, most people have made search engines part of their daily routine. Beyond trivial applications, search engines are increasingly becoming the sole or primary source directing

people to essential information. As internet is growing very fast search engines are playing very important role. Due to the large number of websites, search engines have the complex task of sorting through the billions of pages and displaying only the most relevant pages in the search engine results page (SERP) for the submitted search query.

The paper is organized as follows:

Section 2 contains Information Retrieval Models

Section 3 Literature Review

Section 4 discusses about Search Engines Listing Methods

Section 5 concludes the paper while references are shown in section 6.

II. INFORMATION RETRIEVAL MODEL

Information retrieval has been characterized in a variety of ways, ranging from a description of its goals, to relatively abstract models of its components and processes. Generally, the goal of an information retrieval system is for the user to obtain information from the knowledge resource which helps him/her in problem management. Such functions, or goals, of information retrieval have been described in general models of the type shown in Figure 2.1. This model illustrates basic entities and processes in the information retrieval situation.

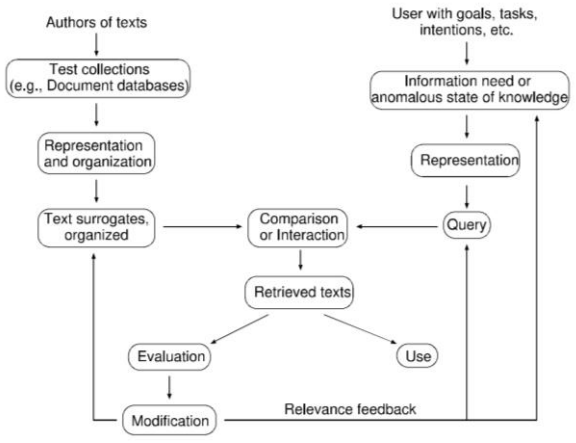

Figure 2.1: A general model of information retrieval.

In this model, a person with some goals and intentions related to, for instance, a work task, finds that these goals cannot be attained because the person's resources or knowledge are somehow inadequate. A characteristic of such a situation is an anomalous state of knowledge or information need, which prompts the person to engage in active information-seeking behaviour, such as submitting a query to an information retrieval system. The query that must be expressed in a language understood by the system is a representation of the information need. This is shown in the right-hand side of Figure 2.1. Due to the inherent difficulty of representing anomalous state of knowledge's, the query in an information retrieval system is always regarded as approximate and imperfect.

On the other side of Figure 2.1, the focus of attention is the information resources that the user of the information retrieval system will eventually access. Here, the model considers the authors of texts; the grouping of texts into collections (e.g., databases); the representation of texts; and the organization of these representations into databases of text surrogates. A typical surrogate would consist of a set of index terms or keywords.

The comparison of a query and surrogates, or in some cases, direct interaction between the user and the texts or surrogates (as in hypertext systems), leads to the selection of possibly relevant retrieved texts. These retrieved texts are then evaluated or used, and either the user will leave the information retrieval system, or the evaluation leads to some modification of the query, the information need, or, more rarely, the surrogates. The process of query modification through user evaluation is known as relevance feedback [6] in information retrieval.

In this section, we show the following three classic information retrieval models:

2.1 Boolean Model: The Boolean model is a simple retrieval model based on set theory and

Boolean algebra. Since the concept of a set is quite intuitive, the Boolean model provides a framework which is easy to grasp by a common user of an information retrieval sys- tem. Furthermore, the queries are specified as Boolean expressions that have precise semantics. Given its inherent simplicity and neat formalism, the Boolean model received great attention in past years and was adopted by many of the early commercial bibliographic systems.

However, the Boolean model has the following two drawbacks. First, its retrieval strategy is based on a binary decision criterion (i.e., a document is predicted to be either relevant or non-relevant) without any notion of a ranking, which prevents good retrieval performance. Second, while Boolean expressions have precise semantics, it is not often simple to translate an information need into a Boolean expression. In fact, most users find it difficult and awkward to express their query requests in terms of Boolean expressions. The Boolean expressions actually formulated by users are often quite simple. Despite these drawbacks, the Boolean model is the standard model for current large scale, operational information retrieval systems.

2.2 Vector Space Model: The vector space model [7] recognizes that the use of binary weights like Boolean model is too limiting and proposes a framework where partial matching is possible. This is accomplished by assigning non-binary weights to index terms in queries and in documents. These term weights are ultimately used to compute similarity between each document stored in the system and the user query. By sorting the retrieved documents in decreasing order of this similarity, the vector model takes into account documents which match the query terms only partially. The main resultant effect is that the ranked document answer set is a lot more precise in the sense it better matches the user information need than the document answer set retrieved by the Boolean model. The vector space model can be summarized as Figure 2.2.

The main advantages of the vector space model are: (1) its term-weighting scheme improves retrieval performance; (2) its partial matching strategy allows retrieval of documents that approximate the query conditions; and (3) its cosine ranking formula sorts the documents according to their similarity to the query. Theoretically, the vector model has the disadvantage that index terms are assumed to be mutually independent.

Despite its simplicity, the vector space model is effective ranking strategy with general collections. It yields ranked answer sets which are difficult to improve on without query expansion or relevance feedback within the framework of the vector space model. Although a large variety of alternative ranking methods has been compared with the vector space model, the consensus seems to be that, in general, the vector space models either superior or almost as good as the known alternatives. Furthermore, it is simple and fast. For these reasons, the vector space model is a popular retrieval model.

Definition *Vector space model*
For the vector space model, the weight $w_{t_i}^d$ is positive and non-binary. Furthermore, the index terms in the query are also weighted. Let q_{t_i} be the weight of query vector $\boldsymbol{Q}$. Then, the query vector $\boldsymbol{Q}$ is defined as $\boldsymbol{Q} = (q_{t_1}, q_{t_2}, \cdots, q_{t_m})$ where m is the total number of index terms in the system. As before, the vector for a document d is represented by $\boldsymbol{w}^d = (w_{t_1}^d, w_{t_2}^d, \cdots, w_{t_m}^d)$.

Let N be the total number of documents in the system and $df(t_i)$ be the number of documents where the index term t_i appears. Let $tf(t_i, d)$ be the raw frequency of term t_i in the document d (i.e., the number of times the term t_i is mentioned in the text of the document d). Then, the normalized frequency $tf_{norm}(t_i, d)$ of term t_i in document d is given by,

$$tf_{norm}(t_i, d) = \frac{tf(t_i, d)}{\sum_{j=1}^{m} tf(t_j, d)}.$$

If the term t_i does not appear in the document d then $tf_{norm}(t_i, d) = 0$. Furthermore, let $idf(t_i)$, inverse document frequency for t_i, be given by,

$$idf(t_i) = \log \frac{N}{df(t_i)}.$$

The best known term-weighting schemes use weights that are given by,

$$w_{t_i}^d = tf_{norm}(t_i, d) \cdot idf(t_i) = \frac{tf(t_i, d)}{\sum_{j=1}^{m} tf(t_j, d)} \cdot \log \frac{N}{df(t_i)},$$

or variation of this formula. The term weighting strategies described above are called TF-IDF schemes.

Figure 2.2: Definition of the vector space

2.3 Probabilistic Model: Robertson et al. [5] Introduced the classic probabilistic model. The model later be- came known as the Binary Independence Retrieval (BIR) model.

The probabilistic model is based on the following fundamental assumption.

Given a user query q, the probabilistic model assigns to each document d, as a measure of its similarity to the query, the ratio, P (d relevant-to q)/P (d non-relevant-to q) which computes the

likelihood of the document d being relevant to the query q. Taking the likelihood of relevance as the rank minimizes the probability of an erroneous judgment [8, 2]. The probabilistic model can be summarized as shown in Figure

2.3

Definition *Probabilistic model*

For the probabilistic model, the index term weight variables are all binary i.e., $w_{t_i}^d \in \{0,1\}$, $w_{q_i} \in \{0,1\}$. A query q is a subset of index terms. Let R be the set of documents known (or initially guessed) to be relavent. Let $\bar{R}$ be the complement of R (i.e., the set of non-relevant documents). Let $P(R|d)$ be the probability that the document d is relevant to the query q and $P(\bar{R}|d)$ be the probability that d is non-relevant to q. The similarity $sim(d,q)$ of the document d to the query q is defined the ratio,

$$sim(d,q) = \frac{P(R|d)}{P(\bar{R}|d)}.$$

Figure 2.3: Definition of the probabilistic model.

III. LITERATURE REVIEW

Our main goal is to improve the quality of web search engines. In 1994, some people believed that a complete search index would make it possible to find anything easily. According to Best of the Web 1994 -- Navigators, "The best navigation service should make it easy to find almost anything on the Web (once all the data is entered)." However, the Web of 1997 is quite different. Anyone who has used a search engine recently can readily testify that the completeness of the index is not the only factor in the quality of search results. "Junk results" often wash out any results that a user is interested in. In fact, as of November 1997, only one of the top four commercial search engines finds itself (returns its own search page in response to its name in the top ten results). One of the main causes of this problem is that the number of documents in the indices has been increasing by many orders of magnitude, but the user's ability to look at documents has not. People are still only willing to look at the first few tens of results. Because of this, as the collection size grows, we need tools that have very high precision (number of relevant documents returned, say in the top tens of results). Indeed, we want our notion of "relevant" to only include the very best documents since there may be tens of thousands of slightly relevant documents. This very high precision is important even at the expense of recall (the total number of relevant documents the system is able to return). There is quite a bit of recent optimism that the use of more hyper textual information can help improve search and other applications [Marchiori 97] [Spertus 97] [Weiss 96] [Kleinberg 98]. In particular, link structure and link text provide a lot of information for making relevance judgments and quality filtering. Google makes use of both link structure and anchor text.

IV. SEARCH ENGINES LISTING

Search engines present two types of listings to users – organic, natural search results and paid search or pay-per-click (PPC) listings – illustrated with the annotated screenshot below in figure 4.1:

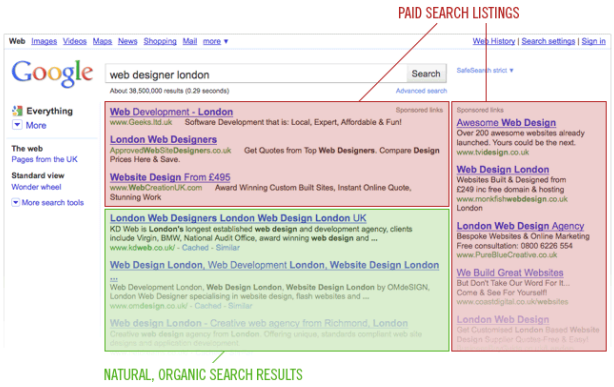

Figure 4.1: Natural Search Results and Paid Search Listing.

The sites listed in each of those two types, and their rank order is determined by very different processes.

4.1 Natural Search Results

A search engine is essentially a very large database containing a record of individual web pages from all over the web. The mechanics behind a search engine can be thought of in terms of three main elements: a search engine spider, a storage database and a relevancy algorithm given below in figure 4.2:

Figure 4.2: Mechanics of Search Engine

The spider or robot is an automated programme which finds web pages and stores a record of them in the search engine's database. A spider is used in this way because of the size of the search engine's database. For example, Google's index currently contains a record of more than 12 billion web pages, and so too manually compile a database of this size would take an unfeasible length of time. These database records are then shown on search engine results pages in the order that is determined by a relevancy algorithm.

Using Google as an example, the following figure 4.3 provides a simple illustration of this process of a search engine accessing, indexing and presenting a list of results in response to a search query:

Each number in the diagram above represents a step in the following process:

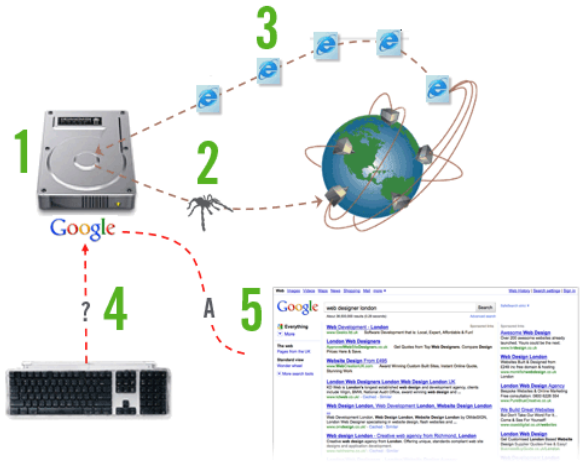

Figure 4.3: How a search engine accesses website content

1. From the central search engine server, the search engine spider (numbered as 2 above) follows links between different websites, thereby 'crawling' the world wide web looking for web pages and documents to index.
2. When a search engine spider finds new content on a website (depicted as 3 in the diagram above), it will attempt to record (index) it within the search engine's database. If there are no barriers present to this process and a web page is indexed successfully, a complicated mathematical algorithm is then applied to the content of the page. This determines what the page is about and how it should be ranked in relation to other documents that have already been stored within the database.
3. When a user types in a search query, Google searches its database for all of the web pages and documents it has indexed and that it considers relevant to that particular search query. It is important to note that any websites that have not been visited by Google's spider and recorded within the database will not be matched to any search query. Therefore, to say you are "searching the web" for something is a somewhat of a misnomer – you are, in fact, searching a database of web pages that Google has previously recorded and formed an opinion about.
4. The resulting web pages and documents that Google finds in its database are returned to the user as listings on a search engine results page. Web pages and documents are ranked on the SERP according to how well they fulfil the criteria set out within Google's latest algorithm.
5. Whilst SEO typically delivers the best return on investment of any marketing channel, it can take some time before you see results as changes made to your site or your link profile need to build to a point where the search engines review your relevancy and credibility before awarding you with higher and broader rankings. On that basis,

depending on your starting position, the competitiveness of your industry and the speed at which your site reflects our recommendations, positive results will begin to materialise from month 2 onwards, and grow incrementally thereafter.

4.2 Paid Search Listings

Paid search marketing, also known as pay per click or PPC, and sometimes (incorrectly) as simply SEM, refers to methods of guaranteeing a website's visibility within search engine results pages by paying a search engine a fee for either strategically placed advertisements or sponsored listings within their search results.

Sponsored listings are mostly given a clear separation from natural, organic listings on search engine results pages. Search engines offering pay per click services allow customers to pay for text ad listings to be shown on searches for specific keywords – essentially the advertiser selects the keywords (search terms) for which they would like their ad to be shown and bids for a position on the relevant search engine results.

By bidding against specific keywords/search terms, listings appear across a network of search engines and search portals such as Google, Bing and Yahoo. When those keyword searches are performed, and your ad is clicked on, you pay the search engines on a per click basis.

The process is however a little more complex. Your position in a list of ads and what you pay per click is the result not only of the maximum cost-per-click you have communicated to the search engines that you are willing to pay, but also what is called you 'Quality Score'. Your Quality Score is fundamentally a score given to you by the search engines that is the product of a calculation of (principally) your click-through-rate (CTR), i.e. the percentage of users who click on your ad over your competitors, and the relevancy of your landing page to the query made by the searcher. By computing this Quality Score, Google can confidently and automatically decide which advertisers should enjoy the highest positions in a list of ads, based on their relevance.

The huge advantage of paid search over most other marketing methods is its immediacy. A paid search campaign can be delivering relevant and transaction-motivated potential customers within a few weeks.

V. CONCLUSION

In order to improve retrieval accuracy of Web search, we studied methods for indexing the contents of Web pages more accurately, and adapting search results according to each user's need for relevant information. Our proposed approaches described in this paper contribute for indexing a target Web page more accurately, and allowing each user to perform more fine-grained search that satisfy his/her information need.

VI. REFERENCES

[1] http://www.w3.org/People/Berners-Lee/WorldWideWeb.html

[2] N. Fuhr. Probabilistic Models in Information Retrieval. The Computer Journal, 35(3): pages 243–255, 1992.

[3] http://www.bloomberg.com/video/67758394

[4] http://en.wikipedia.org/wiki/Web_browser

[5] S. E. Robertson and K. S. Jones. Relevance Weighting of Search Terms. Journal of the American Society for Information Sciences, 27(3): pages 129–146, 1976.

[6] J. Rocchio. Relevance Feedback in Information Retrieval. In G. Salton, editor, The Smart Retrieval System: Experiments in Automatic Document Processing, pages 313–323. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[7] G. Salton. The Smart Retrieval System: Experiments in Automatic Document Processing. Prentice - Hall, Englewood Cliffs, NJ, 1971.

[8] C. J. van Rijsbergen. Information Retrieval. Butterworths.

[9] http://en.wikipedia.org/wiki/Search_engine

[10] http://www.pewinternet.org/Reports/2012/Search-Engine-Use-2012/Summary-of-findings.aspx

[11] http://searchenginewatch.com/article/2049695/Top-Google-Result-Gets-36.4-of-Clicks -Study

[12] http://www.seomoz.org/article/search-ranking-factors

[13] G. Jeh and J. Widom. Scaling Personalized Web Search. In Proc. of the 12th International World Wide Web Conference (WWW 2003), pages 271–279, 2003.

[14] T. Hofmann. Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis. In Proc. of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '03), pages 259–266, 2003.

[15]IBM Almaden Research Centre. Clever Searching http://www.almaden.ibm.com/cs/k53/clever.html.

Author's Profile

Trilok Gupta was born in Kota (Rajasthan). He has done Diploma in Computer Science and received his Master Degree in Computer Science from JRNRV University, Udaipur, Rajasthan-India. He is pursuing Ph.D. Computer Science from Faculty of Computer Application, Pacific University, Udaipur-PAHER, Rajasthan - India. His area of interest includes Data Handling, Data Mining, Web Applications, Search Engines Optimization and Information Exploring. He is working in the field of education for last 14 years. He has published several research papers in National and International Journals. He is currently exploring the anatomy of Search Engines and Web Mining.

Archana Sharma was born in Ajmer (Rajasthan). She is Ph.D in Computer Science and Engineering with specialization in Simulation and Modeling. She completed her M.Tech in Computer Science from Banasthali Vidyapith, India. Her field of study is Simulation and modeling, data mining, database, Artificial Intelligence.

She is working in the field of education for last 15 years. She has taught many subjects at undergraduate and postgraduate level. She has published several research papers in national and International Journals. She is currently working in the field of Cloud Computing, Artificial Intelligence and Educational Data Mining. She is Senior Member of International Association of Computer Science and Information Technology (IACSIT). She is also the Board member in Seventh Sense Research Group Journals. P

Professor Sharma is member of Indian Society of Theoretical and Applied Mechanics. She worked as Editor in Journal of Management and IT 'OORJA'.