Information Filtering and Retrieval: An Overview

Colm O' Riordan
IT Centre,
NUI, Galway,
Ireland.
c.oriordan@ucg.ie

Humphrey Sorensen
Department of Computer Science,
University College Cork,
Ireland.
h.sorensen@cs.ucc.ie

Abstract

The areas of information retrieval(IR) and information filtering(IF) have become very active research domains. The problems created by the large increase of available online information, of which the vast majority is largely unstructured, have accentuated the need for effective mechanisms to separate the relevant information from the irrelevant. This paper reviews the main approaches and systems used in IR and in the newer field of IF. The paper also includes an overview of systems which utilise social or collaborative filtering techniques to deal with the problem of information overload.

1 Introduction

This paper aims to provide an overview of the major developments and approaches in the fields of Information Filtering and Information Retrieval. We look at the advantages and flaws associated with currently existing systems.

The first section outlines the need for such filtering and retrieval systems to overcome the increasingly prevalent problem of information overload.

In section 3, we compare and contrast IR and IF while outlining the main components of retrieval and filtering systems. The later sections discuss the differing approaches.

Section 7 deals with the newer approach of Collaborative filtering. Finally, in the conclusion, we specify the necessary requirements of any filtering/retrieval system and outline our work in this regard.

2 Information Overload

The ever-increasing popularity of the Internet has led to an influx of users and, consequently, a huge increase in the volume of available on-line data. Such data is available through web sites, ftp sites, mailing lists and USENET newsgroups. This increase has led to a situation where users are swamped with information and have difficulty sifting through the reams of material, much of which is not relevant to them. This scenario is commonly referred to as the problem of information overload.

In an effort to protect the user from the problem of information overload, numerous mechanisms have been developed to aid him/her in separating relevant from irrelevant information. These approaches have included:

- **Retrieval/Filtering Systems**: These systems deal with the ranking of semi-structured or unstructured data (usually textual) in order of relevance. Retrieval refers to the selection of data from a fixed data (or document) set, whereas filtering typically refers to selection of relevant information (or rejection of irrelevant information) from a stream of incoming data. Retrieval systems are generally concerned with satisfying a user's one-off information need (query); filtering systems are usually applied to attaining information for a user's long term interests (profiles).
- **Hyper-textual links**: With the increase in information available in the World Wide Web (WWW), information is commonly organized using the hypertext paradigm where users can retrieve information in a stepwise manner by selecting hyper-text links to other information. This approach to data organisation is best exemplified by the use of HyperText Markup Language (HTML) [19] for document creation and linking.
- **Categorisation**: Data can be categorised into numerous categories or domains of interest. Users can access these categories (often with a hierarchical structure, for example, Yahoo [11]) to attain information relevant to their particular need.

3 IR and IF

Information retrieval (IR) is a well established field in information science, which addresses the problems associated with retrieval of documents from a collection in response to user queries. Information filtering (IF) is a more recent specialisation within information science, having come to the fore due to increasing volumes of online transient data. Similarities are dissimilarities between IR and IF have been well debated[4] and a relatively coherent viewpoint has emerged.

The primary dissimilarities relate to the nature of the data set and the nature of the user need. Traditionally in IR the data-set is viewed as being relatively static; in IF it is seen as a dynamic 'stream' of information. In IR, user queries represent a one-off short-term information need; in IF, user queries (or profiles) are representative of long-term (possibly dynamic) information needs. This distinction is becoming more blurred with the application of IR/IF to the web - some sites (collections) are relatively static, others very dynamic. The IR/IF research community has been debating other distinctions between the fields of study. Issues here include: (1) the comparative importance of deep/shallow representation (and effort required for same) in each case; (2) diverse methods of document ranking, appropriate to the task at hand; (3) differences in the method of accuracy assessment between IR and IF.

However, [4] focussed to a greater extent on common features of IR and IF systems. In particular, the authors identified common approaches, functionality and components within IR and IF systems. A brief overview of these main components of IR/IF systems is presented in this section. A diagrammatic representation of an IR/IF system is given in Figure 1.

The chief mechanisms arising in IR/IF systems are as follows:

- **Representation:** Both the user's information need (query or profile) and the document set (fixed or dynamic) must be represented in a manner such that the computer can effect comparison. Representation techniques range from using indexes, vector representation or matrices, to the more modern representations—neural networks, connectionist networks and semantic networks.

- **Comparison:** The comparison mechanisms used are dependent on the underlying representations used in the system. The initial primitive systems involve utilising simple string matching algorithms; more advanced systems use statistical operations involving vector and matrix calculations; and some contemporary systems use manipulation of the network (semantic/neural) via spreading activation search mechanisms or propagation techniques.

- **Feedback:** To improve the performance of the IR/IF system, a feedback mechanism is usually incorporated. This usually involves the user stating his/her satisfaction or dissatisfaction with returned documents. On receiving this feedback, the query is usually modified to attain better results and the filtering process begins again. This process can be automatic or manual. Searching the web using many of the common search engines involves a manual modification – by adding more terms etc. whereas other systems involve automatic expansion/modification of profile/query using thesaurai, etc. Again, a more complete overview of feedback mechanisms is provided in a later section.

3.1 Metrics:

The main metrics used to test the accuracy of the retrieval/ filtering algorithm are precision and recall. These are defined as:

$$\text{Precision} = \frac{\text{Number of relevant items retrieved}}{\text{Number of items retrieved}}$$

$$\text{Recall} = \frac{\text{Number of relevant items retrieved}}{\text{Number of relevant items in database}}$$

Typically, the precision decreases as recall increases and vice-versa. A precision-recall graph is usually generated to graphically represent the relationship between the two measures.

4 Information Retrieval Methods

In this section, an overview of previous approaches to information retrieval is given. This is not intended to be a complete or exhaustive overview—but instead tries to focus on the more important systems and approaches.

4.1 String Matching

The simplest approach to information filtering and retrieval is the string matching approach. If a document contains a user-specified string, that document is deemed to be relevant; otherwise, it is deemed irrelevant. This simple approach is very easy to use (and to implement) but is fundamentally flawed. It is based on the assumption that

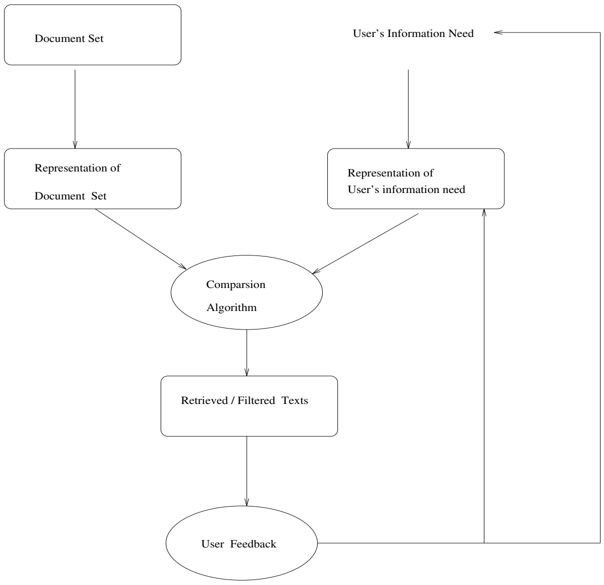

Figure 1. Architecture of a retrieval/filtering system

it is a "simple matter for users to foresee the exact words and phrases that will be used in the documents they will find useful and only in those documents"[6]. Such a system makes no attempt to overcome the problems of *homonymy* - the ability of a word to have a different meaning in different contexts (e.g. gravity), and *synonymy* - numerous words which have the same meaning (e.g. thesis and dissertation).

4.2 Boolean Information Retrieval

This is an extended version of the simple string matching approach where users may filter using a set of words/terms combined via the Boolean primitives **AND**, **OR** and **NOT**. While allowing a more powerful expressive representation of the user's information need, this approach has major drawbacks:

1. This type of input requires some skill on behalf of the user for any reasonably complex information need.
2. Differing vocabularies will result in the same problems that are associated with string matching techniques.

Most commercial systems and web-based systems are based on the Boolean approach, sometimes augmented with proximity operators. Sample systems offering Boolean queries include AltaVista[1], Excite[2], Lycos[3], Inktomi[4], (for an overview of these systems and others, see [17]) and SIFT [44].

[1] Available at http://www.altavista.digital.com
[2] Available at http://www.excite.com
[3] Available at http://www.lycos.com
[4] Available at http:/www.inktomi.berkely.edu

4.3 Vector-Space Model

The vector space model [37] is based on the statistical occurrence of terms in the user profile and the documents (incoming articles). The documents are identified by terms. A document D is represented as a vector of dimension m, where m is the total number of terms used to identify content. The importance of each of these m terms is represented by an associated weight. $D = (w_1, ..., w_m)$, where w_i is the weight assigned to the i-th term, is written to represent a document of m terms.

The assigned weight for each term is used as a measure of its importance. Originally in this model, the value for a term was 0 or 1, representing irrelevance and relevance respectively. It was later expanded to include the notion of a generalised Boolean algebra [36]. Typically words that occur a few times are given a high weight value and are said to have a *high resolving power*, whereas words that occur with a high frequency are given a low value and are said to have a *low resolving power*. Salton uses the following formula for assigning the values:

$$w_t = f_{it}\ log \frac{N}{N_t}$$

where N is the number of documents in the collection, N_t is the number of documents in the collection containing the term t and f_{it} is the number of times t occur in document i.

A query or profile (representing the user's information need) is also represented as a vector. Given a query of m terms, a vector $P(t_1, ..., t_m)$ is used to represent the user's profile.

The most commonly used similarity function is the cosine of the angle between the query vector and the document (article) vector

$$cos(P, D) = \frac{< P.D >}{||P||\ ||D||}$$

where

$<.>$ is the dot product and $||P|| = \sqrt{< P.P >}$

The advantages of this approach are adaptability, robustness and minimal user intervention. While definitely an improvement on the boolean model, the vector space model also suffers from the problems of synonymy and homonymy.

4.4 Use of Thesaurai

A thesaurus is a set of terms (words and phrases) and a set of relations between these terms. Thesaurai have been used to attempt overcome the problems of synonymy and homonymy by expanding the initial query or profile. A thesaurus may be manually constructed (e.g. Roget's Thesaurus) or might be automatically generated, usually based on document collection statistics. Furthermore, it may involve simple word-word substitution or a more sophisticated generalisation-specialisation hierarchy (e.g. WordNet [27]). An automatic thesaurus is usually built on term co-occurrence information. Salton showed that using a synonym thesaurus with relevance feedback produces significant improvement[35].

Applying thesaurai has proved beneficial in attempting to overcome the vocabulary problem, but has the disadvantage that the additional context provided by associated terms in a profile is ignored.

4.5 Inference Networks

The use of inference networks for information retrieval has been explored by Turtle and Croft [43]. A Bayesian inference network is an acyclic dependency graph in which nodes represent dependency relations between propositions. Given probabilities of the root nodes in this graph, probabilities for all remaining nodes may be calculated. The use of these networks for information retrieval represents an extension of the older probabilistic models and allows for the integration of several sources of information.

4.6 Latent Semantic Indexing

LSI attempts to overcome the problems associated with word-based methods, especially the vector space approach, by organising textual information into a semantic/conceptual structure more suitable to information retrieval. The phrase "latent semantic" refers to the inherent underlying associations between words used to express a particular concept. Typically, it is not the terms occurring in a document that are used as a feature set but, rather, the domains in which they occur.

The documents, as in the vector space model, may be viewed as a set of vectors. The set of vectors, or matrix (term by document), is decomposed using a singular value decomposition to an approximation of the original matrix. This approximation matrix is smaller, leading to more efficient comparisons, and takes into account latent interconnections between terms within documents. The user-profile/query is represented in a similar manner. The relevance of a document/article to a profile is determined by taking the dot product between the document and profile to obtain the cosine of the angle between the vectors.

The advantages of LSI for information filtering/retrieval are increased efficiency over methods such as the vector-space model due to reduced vector sizes, and filtering at a level above that of lexical level. The main disadvantage of this approach is the difficulty in determining the number of dimensions k; too many dimensions and the system reduces to the vector space model; too few and loss of context arises resulting in coarse-grained filtering.

Liddy, Paik & Yu[25] used a form of Latent Semantic Indexing by pre-defining a set of domains (termed *Subject Field Codes*). Each term in a document is assigned a set of SFCs. These vectors of SFCs are then used to represent documents and to effect comparisons. Term co-occurrence data is used in overcoming any ambiguity that may exist in assigning the appropriate SFC. (For example, the term *earth* may be assigned the SFCs - ENG, EARTH, ASTR etc., this set might be further reduced based on the occurrence of other terms in close proximity to *earth*).

4.7 Connectionist Approaches

Connectionist networks assume "that information processing takes place through the interactions of a large number of simple processing elements called units, each sending excitory or inhibitory signals to other units" [34]. In most connectionist approaches to information retrieval, each node is used to represent an individual keyword. This is known as a 'local' representation, as opposed to the 'distributed' representation found in other connectionist systems such as the Boltzmann machine [1]. The search mechanism usually used in these systems is the *spreading activation search* (SAS). In this search strategy, activity is propagated throughout the system and nodes with a high level of activity are returned as the result of the search. In using SAS, an attempt is made to find connections between terms that are not directly linked but which are linked via a number of co-occurrence links. This attempt to pay attention to a term's surrounding context, along with the ability to learn, are the main strengths of the approach.

Semantic networks[29], coupled with spreading activation techniques, have been applied to the area of IR. Such systems include the GRANT system [23] and I^3R [14]. [13] provides a thorough review of the application of spreading activation methods and semantic networks to IR. An example of an IR system adopting the connectionist approach is Belew's Adaptive Information Retrieval (AIR) system [2]

[3]. A query is initiated by setting the activity level of the nodes in the connectionist network to a specific level. The activity is then 'leaked' out to the network. In Belew's system the activity is leaked out in a small number of cycles and only small portions of the network ever become active. This differs from other connectionist systems [24], [28], and accounts for the increased speed in Belew's system.

The system has been shown to learn, from relevancy judgments, advanced semantic features such as word-stems, synonyms and simple phrases. This simple algorithm is also capable of learning transitive relations between terms. Other approaches using a connectionist or distributed representation also exist—the interested reader is directed to [15].

5 Filtering Systems

The concepts and approaches detailed in the previous section have been applied to develop a number of retrieval systems. In this section we enumerate some of these systems, with particular emphasis on those which have been adapted and applied to the task of information filtering.

- INQUERY[5] **Callan, Croft [10]** This system, originally applied to information retrieval has also been applied to the information filtering domain [7]. The system is based on a probabilistic model. Bayesian inference networks are used to represent documents and user profiles. Document comparison is achieved using a recursive inference to propagate values through the inference network and then retrieving documents that have the highest ranking.

- SMART[6] **Buckley[8]:** This retrieval system was developed in 1985 - it uses the vector space model with iterative query refinement to focus precision and recall. The SMART system has also been applied to the task of filtering[9].

- SIFT[7] **Yan & Garcia-Molina [44]:** This retrieval system allows users to submit their profiles, representing their long-term information need, via a WWW browser. The user's profiles are then compared to news articles, and those articles deemed relevant are then displayed to the user. Two filtering/retrieval techniques are offered to the user - Boolean keyword match and a model based on the vector space approach.

Other filtering systems (based on concepts from the IR domain) have also been developed and tested against document sets. These include the system developed by Strzalkowski *et al.* [41] which uses natural language processing techniques, Okapi [31] which uses a probabilistic model, the WIN system [42] which utilises inference networks, and a LSI-based system developed by Dumais [16].

6 Relevance Feedback

Relevance feedback has proved to be highly effective for improving information filtering and retrieval. Upon receiving returned articles, the user may provide relevance judgments for these articles. These relevance judgments may subsequently be used to guide the matching function for the retrieval/filtering system.

Typically, on presentation of the results from the filtering system, the user is asked to identify which documents are relevant and which are non-relevant. This information, along with the current user profile P_k, is then used to form a new profile $\mathrm{P}_{(k+1)}$ which is used as the user's profile in future filtering. The feedback mechanism adopted is clearly dependent on the representation (and comparison) strategies in use, but generally involves adding/removing profile terms and adjusting any weighting of existing terms.

6.1 Feedback Techniques

Relevance feedback techniques for two of the most popular filtering methods are:

- Vector-Space Model:
  The Rocchio feedback model [33] is the most common method used. Rocchio showed that a more effective profile representation could be iteratively generated as follows:

  $$\text{Rocchio:} \quad P_{k+1} = P_k + \beta \sum_{k=1}^{n_1} \frac{R_k}{n_1} - \gamma \sum_{k=1}^{n_2} \frac{S_k}{n_2}$$

  where P_{K+1} is the new profile, P_k is the old profile, R_k is a vector representation of a relevant article k, S_k is a vector representation for non–relevant article k, n_1 is the number of relevant documents and n_2 is the number of non–relevant documents. The values β and γ determine the relative contributions of positive and negative feedback, respectively.

  Relevance feedback using this technique has been shown to result in a significant improvement in retrieval performance[36].

- Probabilistic networks:
  A query can be modified by adding the first m terms

[5] Available at http://ciir.cs.umass.edu/inquerypage.html
[6] Available at ftp://ftp.cornell.cs.edu/pub/smart/smart.1..0.tar.Z
[7] Available at http://sift.stanford.edu/

in a list where all terms present in documents deemed relevant are ranked according to the formula[32]:

$$w_i = log\left[\frac{r_i(N - R - n_i + r_i)}{(R - r_i)(n_i - r_i)}\right]$$

where N is the number of documents retrieved, n_i is the number of documents with an occurrence of term i, R is the number of documents deemed relevant by the user, r_i is the number of relevant documents containing an occurrence of term i.

- Feedback techniques for newer filtering models, specifically connectionist, are discussed in [5].

Other feedback issues also exist. For example, passage level feedback involves the user selecting a section of a document as relevant (as opposed to the more usual selection or rejection of a full document). The selected passage is then used as positive feedback, leading to possible incorporation of new search terms or possible re-weighting of existing terms.

Not all feedback mechanisms cause a modification to the user profile. Other possibilities also exist, such as the re-ranking of returned documents as implemented in Browse [22], designed and developed by Higuichi and Jennings. In their system, propagation of activity through neural network representations of the documents, results in a re-ranking of the initially returned documents.

7 Collaborative Filtering

Different criteria may be used to filter documents/articles - the filtering and retrieval systems mentioned so far in this paper all use the content of the documents as the basis for the filtering. Malone [26] describes three categories of filtering technique — cognitive, social and economic. Cognitive filtering, as heretofore discussed, is based solely on the content of the articles. Social filtering techniques are based on the relationships between people and on their subjective judgments. Inserting a certain person's (the sender's) name in a kill file is a primitive form of social filtering, indicating the wish to filter out all information coming from that person. Economic filtering bases filtering on the cost of producing and reading articles. For example, a USENET News filtering system may filter out articles that have been cross-posted to many groups.

Collaborative filtering is a form of social filtering - it is based on the subjective evaluations of other readers attached as annotations to shared documents. Schemes using collaborative filtering use human judgments, which do not suffer from the problems which automatic techniques have with natural language such as synonymy, polysemy and homonymy. Other language constructs, at a pragmatic level, like sarcasm, humour and irony may also be recognised.

7.1 Collaborative Filtering Systems

7.1.1 Tapestry

The Tapestry system [18] was developed to aid users in the management of incoming news articles or mails. It was developed specifically for people working in work-groups. Users can use content filters to select articles and can also retrieve/filter articles based on the recommendations of other users. On reading articles, users are asked to provide a rating of the article.

As Shardanand and Maes [38] point out, in order to receive recommended articles, users must know in advance the names of the authors who have previously recommended the articles, i.e, "the social information filtering is still left to the user".

7.1.2 GroupLens

GroupLens[8] is a "distributed system for gathering, disseminating, and using ratings from some users to predict other user's interests in articles" [30]. It is based on the premise that users within a group known to have similar interests will have similar interests in future.

Ratings are submitted by users upon reading an article posted to a newsgroup. A rating scale of 1-5 exists, a score of 1 indicating the article is not relevant or not worth reading, while a score of 5 signifies the article is relevant and worthy of reading. The users' ratings are distributed via the USENET propagation scheme. The user ratings are posted to newsgroups dedicated to the posting of GroupLens ratings.

The scoring method used is based upon the heuristic that people who agreed in the past are likely to agree again in the future, particularly if the article is from the same newsgroup. The main difficulties with GroupLens are the limited number of newsgroups catered for and the time required by a user to rank all articles read for this system to be effective.

7.1.3 Recommender Systems

Other common examples of social or collaborative filtering include recommender systems. In these systems, users rate different interests, such as, videos (e.g *Bellcore's video*

[8] Available at http://www.cs.umn.edu/Research/GroupLens/

recommendation [20] and musicians (e.g. in Firefly[9], previously known as Ringo [38]). Films or musicians are then recommended to the user based on comparisons with other users' rankings. The recommendation of new films or artists is also based on the premise that people with similar interests in the past will have the same interests or likes in the future.

7.1.4 Beehive

Beehive [21] is a distributed system for social filtering of information. The system consists of two main parts:

- Information Gathering: This section records dealings with other users in the community (from email header files, etc.). A table of e-mail addresses of people with similar interests to the user is created. An entry in the table consists of a community field (e.g. friends, researchers, etc.), a threshold value and a list of email addresses.
- Information Dissemination: This section allow the user to drop text (a document, a URL etc.) onto an icon, the information then being disseminated to all people in the user's community file.

Two advantages of this type of filtering over other recommender systems are (i) its inherent distributive nature avoids problems associated with central databases and (ii) its domain is not limited.

7.2 Observations

Social or collaborative filtering addresses issues ignored by simple cognitive systems which have predominated to date. They have been influenced partly by existing IR/IF theories and partly by user modeling concepts - e.g. the recognition that like-minded individuals frequently co-operate to attain shared goals. The large quantities of on-line information can clearly be rendered more manageable via word-of-mouth recommendations among cooperating consumers.

However, existing systems either promote collaboration within a limited domain (as in Firefly of LifeStyleFinder) or require explicit user intervention (as in GroupLens where, for optimal results, all users must score articles in the range 1-5, with reasonable consistency). For a collaborative filter to be most beneficial it must fulfill the following requirements:

1. It must filter articles with high precision and recall
2. It should promote cooperation with other users over a large domain
3. It should be unobtrusive in it's operation.

[9] Available at http://www.firefly.com/

8 Conclusion

In this paper we have looked at some existing filtering and collaborative filtering systems. We also looked at the theory behind the approaches used in these systems and their respective strengths and weaknesses.

Our work in information filtering has tried to overcome some of the weaknesses associated with previous attempts:

1. Our information representation utilises a semantic network for document representation and a spreading activation search methodology to effect comparisons. This has been shown to be effective[39].
2. We incorporate a feedback mechanism which modifies the existing graph to reflect more accurately the user's information need.
3. We have utilised a multi-agent system (communicating via the Contract Net Protocol) to allow intelligent collaborative filtering over a large range domains.
4. We have investigated the incorporation a web-searching facility, which uses existing web-indexes in conjunction with a more intelligent filtering mechanism to locate and retrieve web-based information.

More detailed information on our work and results is available in [12], [40], and [39].

References

[1] D. Ackley, G. Hinton, and T. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 9, 1985.

[2] R. K. Belew. Adaptive information retrieval: Machine learning in associative networks. 1986.

[3] R. K. Belew. Adaptive information retrieval: Using a connectionist representation to retrieve and learn about documents. *Proceedings of the Twelfth Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, 1989.

[4] N. Belkin and B. Croft. Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(2), December 1992.

[5] P. Biron and D.Kraft. New methods for relevance feedback: Improving information retreival performance. 1993.

[6] D. Blair and M. Maron. An evaluation of retrieval effectiveness for a full document retrieval system. *Communications of the ACM*, 28(3), March 1985.

[7] J. Broglio, J. Callan, B. Croft, and D. Nachbar. Document retrieval and routing using the inquery system. November 1994.

[8] C. Buckley. Implementation of the smart information retrieval system. Technical report, Dept. of Computer Science, Cornell University, 1985.

[9] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using smart: Trec 3. November 1994.

[10] J. P. Callan, W. B. Croft, and S. M. Harding. The inquery retrieval system. Technical report, Department of Computer Science, University of Massachusetts, 1993.

[11] A. Callery. Yahoo!, cataloging the web [online], available at http://www.library.ucsb.edu/untangle/callery.html. *Proceedings of the Conference sponsored by the Librarians Association of the University of California*, 1996.

[12] C.O'Riordan, H. Sorensen, and A.O'Riordan. Multi-agent collaborative filtering. *Workshop on Social Agents on the Web, ECSCW'97, Available online at: http://orgwis.gmd.de/projects/SAW/ORiordan.html*, 1997.

[13] F. Crestani. Application of spreading activcation techniques in information retrieval. *Artificial Intelligence Review*, pages 453 – 483, 1997.

[14] W. Croft and R. Thompson. I^3r: a new approach to the design of document retrieval sy stems. *Journal of the American Society for Information Science*, 6(38):389 – 404, 1987.

[15] T. E. Doszkocs, J. Reggia, and X. Lin. Connectionist models and information retrieval. *Annual Review of Information Science and Technology*, 25:209–260, 1990.

[16] S. Dumais. Latent semantic indexing(lsi) routing for trec-3. November 1994.

[17] A. Eagan and L. Bender. Spiders and worms and crawlers, oh my: Searching on the world wide web. *Proceedings of the Conference sponsored by the Librarians Association of the University of California*, 1996.

[18] D. Goldberg, D. Nichols, B. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61 – 70, December 1992.

[19] M. J. Hannah. Html reference guide. Technical report, Sandia National Laboraties, 1996.

[20] W. Hill, L. Stead, M. Rosenstein, and G. Furnas. Recommending and evaluating choices in a virtual community of use. 1994.

[21] B. Huberman and M. Kaminsky. *Beehive* : A system for cooperative filtering and sharing of information. *Computer Human Interaction*, pages 210–217, 1996.

[22] A. Jennings and H. Higuchi. A personal news service based on a user model neural network. *In IEICE Transactions on Information and Systems*, 1992.

[23] R. Kjelden and P. Cohen. The evolution and performance of the grant system. *IEEE Expert*, pages 73 – 79, 1987.

[24] K. L. Kwok. A neural network for probabilistic information retrieval. *Proceedings of the Twelfth Annual International ACMSIGIR Conference on Research and Development in Information Retrieval*, 1989.

[25] E. Liddy, W. Paik, and E. Yu. Text categorisation for multiple users based on semantic features fr om a machine-readable dictionary. *ACM Transactions on Information Systems*, 12(3):278–295, July 1994.

[26] Malone, Grant, Turbak, Brobst, and Cohen. Intelligent information sharing systems. *Communications of the ACM*, 30(5):390–402, 1987.

[27] G. Miller. Wordnet: An online lexical database. *International Journal of Lexicography*, 3(4):235 – 312, 1996.

[28] M. C. Mozer. Inductive information retrieval using parallel distributed computation. Technical report, Institute for Cognitive Science, 1984.

[29] R. Quillian. Semantic memory. In M. Minsky, editor, *Semantic Information Processing*, pages 216 – 270. MIT Press, 1968.

[30] P. Resnick, N.Iacovou, M. Suchak, P. Bergstrom, and J. Reidl. Grouplens : An open architecture for collaborative filtering of netnews. *Proceedings of ACM 1994 Conference on CSCW*, pages 175 – 186, 1994.

[31] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. November 1994.

[32] S. F. Robertson. The probability ranking principle in ir. *Journal Of Documentation*, pages 294–304, 1977.

[33] J. Rocchio. Relevance feedback in information retreival. pages 313–323, 1971.

[34] D. E. Rumelhart and J. L. McClelland. *Parallel Distributed Processing*, volume 1. MIT Press, 1986.

[35] G. Salton. *Automatic Information Organization and Retrieval*. McGraw-Hill Book Company, 1968.

[36] G. Salton. *Automatic Text Processing: The transformation, Analysis, and Retreival of Information by Computer*. Addison-Wesley, 1989.

[37] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw Hill International, 1983.

[38] U. Shardanand and P. Maes. Social information filtering: Algorithms for automating "word of mouth". 1995.

[39] H. Sorensen, A. O'Riordan, and C. O'Riordan. Personal profiling with the informer filtering agent. *Journal Of Universal Computer Science*, 3(8), 1996.

[40] H. Sorensen, A. O'Riordan, and C. O'Riordan. Text filtering with the informer interface agent. *EXPERSYS-96*, 1996.

[41] T. Strzalkowski, J. Carballo, and M. Marinescu. Automatic query expansion using smart: Trec 3. November 1994.

[42] P. Thompson and H. Turtle. Automatic query expansion using smart: Trec 3. November 1994.

[43] H. Turtle and W. Croft. Evaluation of an inference network-based retrieval model. *ACM Trans. on Info. Systems*, 3, 1991.

[44] T. Y. Yan and H. Garcia-Molina. Sift - a tool for wide-area information dissemination. pages 177–186, 1995.