International Journal of Hybrid Information Technology
Vol.6, No.6 (2013), pp.345-358
http://dx.doi.org/10.14257/ijhit.2013.6.6.31

An Analytical Framework for Web Information Filtering Techniques

Narges Sadat Khozooii[1], Saman Haratizadeh[2] and Mohammad Reza Keyvanpour[1]

[1]*Department of Computer Engineering AlZahra University Tehran, Iran*
[2]*Faculty of new sciences and technologies, university of Tehran, Iran*

nkhozooii@gmail.com, haratizadeh@ut.ac.ir, keyvanpour@alzahra.ac.ir

Abstract

Internet is a huge source of information. The growth of information and unstructured and semi structured nature of web information, cause some challenges for people to find their potential useful information for long-time needs. Hence, the implementation of automated tools selecting and evaluating information is necessary. Information filtering could be seen as a solution to this problem. It allows us to automatically filter out the unwanted content of the information. In this paper we first define the analytical architecture of the web information filtering system and second we suggest a systematic framework to classify web information filtering structures. We hope this proposed framework will lead to empirical and technical comparison of web information filtering structures and development of more efficient structures at future.

Keywords: *Information filtering, Information retrieval*

1. Introduction

Todays, volume of information available on the internet increases. Users can be easily overloaded with this information. Thus it is necessary for users to access to the most interesting and valuable documents quickly and in a limited time. One solution of this problem is filtering. This task, firstly, introduced by Luhn in 1958 as "Selective Dissemination of Information" and named "filtering" by Denning in 1975. An Information filtering (IF) [1] system monitors an incoming document stream to find the documents that match information needs of users. IF systems:

- Are applicable for unstructured or semi-structured data (*e.g.*, documents, e-mail messages);
- Handle large amounts of data;
- Deal primarily with textual data;
- Are based on user profiles; and
- Their objective is to remove irrelevant data from incoming streams of data items [2].

We can view IF as a special type of Information retrieval (IR). Many techniques in IR is common with IF but IF has some characteristics that separate it from IR. IF systems are designed for long term users that their long term needs, are designed in user profiles. The goal of filtering process is to imply the removing data from an incoming stream, instead of retrieving. Each of filtering and retrieval systems has three

ISSN: 1738-9968 IJHIT

components. (i) Document representation. (ii) User's interests representation. (iii) Algorithms used to match user's interest to documents representation [3].

Information filtering systems make use of techniques from two research areas, information retrieval and user modeling. There is three kind of filtering: Cognitive, Social and Economic filtering. Cognitive filtering uses the content of incoming documents and the information needs of a user are used to intelligently match messages to receivers; this is what is now known as content-based filtering.

Social filtering supports the personal and organizational inter-relationships of individuals. This approach complements the cognitive approach by judging the potential of a message based not only on its representation but also on the characteristics of its sender and other users; this is now commonly referred to as collaborative filtering.

Economic filtering involves the use of various kinds of cost-benefit assessments with explicit or implicit pricing mechanisms are used to guide the document filtering process [4].

The organization of this paper is as follows: in second part, we define the analytical architecture of the web information filtering system and explain each of these architecture components. In third part we propose a systematic framework to classify web information filtering structures. Finally, we close the paper with our conclusions.

2. The Architecture of the Web Information Filtering Systems

Here we consider to the architecture of information filtering systems. Note that we have more centralization on textual information filtering. We can see The Architecture of the Web Information Filtering Systems in Figure 1.

2.1. Information repository

This unit is responsible for gathering dynamic information and analysis them to a credible format for filtering. Information can be documents, images, sounds or videos. This unit usually uses the result of search engines for collecting information. Search engines integrate a big number of sources including web search engines, domain specific portals, catalogues / directories and databases. This unit contains interesting and uninteresting information together.

2.2. User Profile

User interests are modeled via user profiles. User profiles can be defined for an individual user or for a group of users. User profiling can be studied under various aspects: Classification of users' profiles, Associated generic models, Contents representation models and Defining profiles methods [5]. Not that the four children of the root "User Profiles", at Figure 2, do not represent four distinct classes but actually only four different points of view for classification of user profiles.

2.2.1. Classification of users' profiles: As you see in Figure 2, we can classify users' profiles in four classes: Users' needs profiles, Users' judgments profiles, Users' demographic data profiles, Multi-criteria profiles.

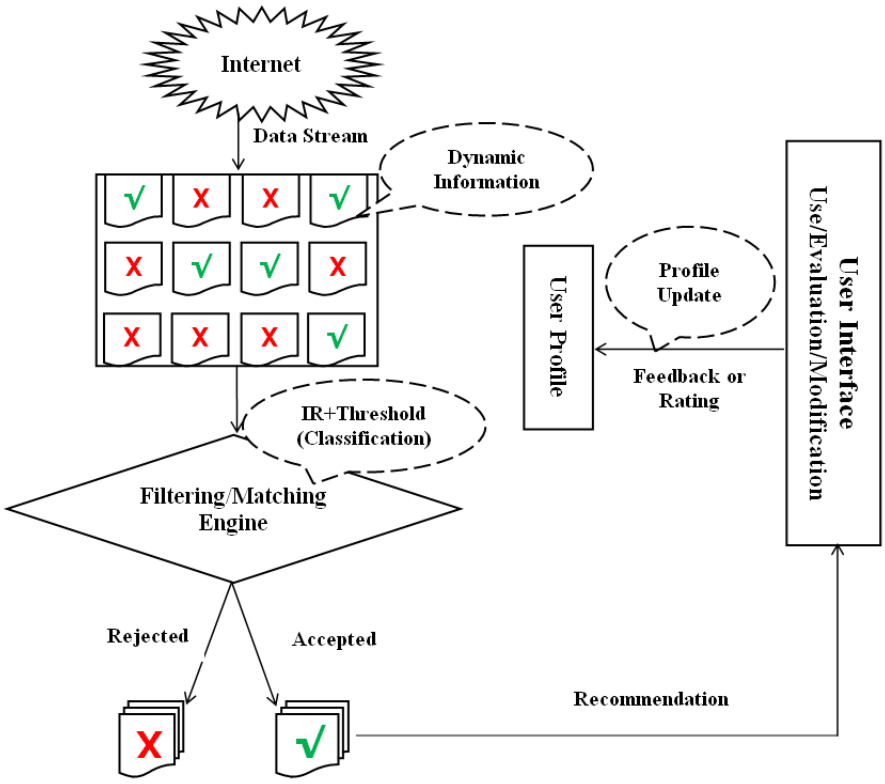

Figure 1. The Architecture of the Web Information Filtering Systems

- **Users' needs profiles:** These profiles describing the users' needs or interest centers.
- **Users' judgments profiles:** These kinds of profiles contain judgments of users on a set of documents.
- **Users' demographic data profiles:** Profiles describing different demographic data of the users: name, gender, age, profession, address, and so on.
- **Multi-criteria profiles:** Profiles describing various characteristics of the users: needs, judgments, demographic data, and so on.

2.2.2. Associated generic models: The description of a user profile generally follows a given model. Two widely used models for users' profiles representation are: the attribute-value model and the hierarchical model.

- **Attribute-value model:** The attribute-value model describes a user profile by a set of independent attributes bounded to an atomic value (string, numeric, date…). For instance, one may describe a user's demographic data with the attribute value model as follows: (name, Peter), (gender, male), (age, 18), (job, student). The first element of each previous pair represents the attribute and the second element describes the value bounded to this attribute. The attribute is considered as a key so in an attribute-value model two attributes with the same name cannot exist at the same time.

- **The hierarchical model:** The hierarchical model organizes various characteristics (or attributes) of a user profile as a tree where leaves are bounded to values representing the contents of the profile. The hierarchy is a way of defining a relationship between the different characteristics. Hence, each non-leaf node represents a class of attributes. Moreover, in a hierarchical model two attributes with the same name but with different access paths in the hierarchy can exist at the same time.

2.2.3. Contents representation models: Generally, contents of users' profiles are represented by a list of atomic values or by a list of weighted values.

- **Atomic values:** Atomic values lead to a form of database matching when comparing different attributes while weighted values allow a fuzzier matching that evaluates a degree of similarity between different attributes instead of a binary similarity value.
- **Weighted values:** Weighted values are mostly used in IR and the cosine formula is generally used to measure inter-profile or query-profile similarities [6].

2.2.4. Defining profiles methods: Manual methods are often used to define profiles; to do that a user generally fills in a form. On the other hand, automatic or semi-automatic methods may also be used to define attribute contents of profiles such as indexing, clustering, and profiling and stereotypes approaches:

- **Indexing** consists in selecting the keywords that best characterize a text (document, query ...). For each keyword a weight is calculated by using tf-.idf like formulas [6]. So one may define the given user's interests by indexing the set of documents that he has visited, saved or judged [7].
- **Clustering or machine learning** [8] consists in identifying objects classes based on similarity of their characteristics. Clustering tries to minimize variance inside a given class and to maximize this variance between classes. Clustering result is then a set of heterogeneous classes with homogenous contents. Hence, one may create users' profiles by applying clustering methods on the set of document contents they saved, judged or visited in order to discover a user's interests or topics.
- **Stereotypes** [9] consist in pre-defining classes and characteristics of these classes. Documents or users are automatically bounded to a given classes according to their contents. The stereotypes approach is a kind of clustering and is mainly used for defining users' groups.
- **Profiling** [10] consists in tracking the user during his different log sessions and in analyzing his behavior. Profiling helps to find documents saved or judged by a user. Therefore profiling is mostly used in electronic commerce in order to identify which kinds of products a user is looking for and then recommend him items that meet these needs. For that purpose, profiling generally analyses clicks on products, products saved in a shopping basket, purchase of products…

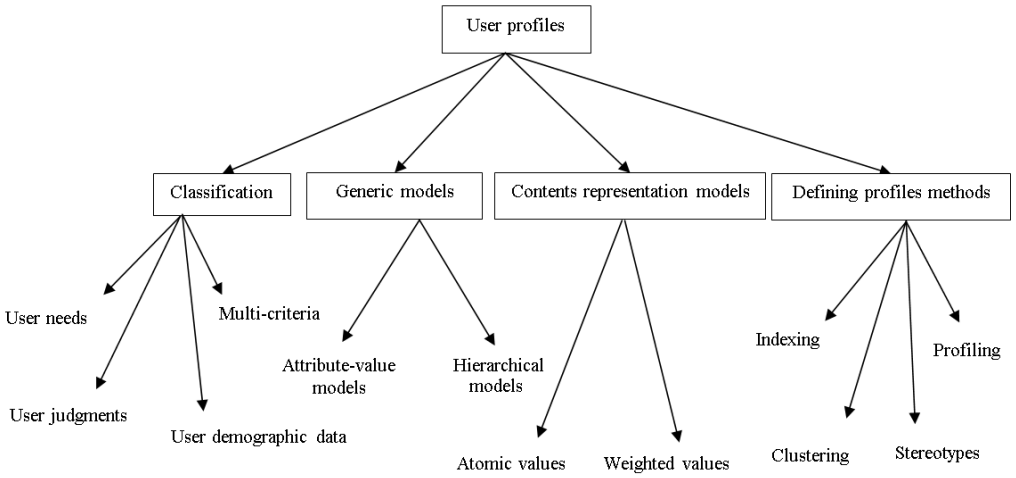

Figure 2. User Profiles Analysis

2.3. Filtering / Matching Engine

This unit compares user profile with the collected data items. It mark accepted to data item if it is an interesting item (match with profile) and mark rejected if the data item is not an interesting item(don't match with profile). Accepted data send to the user interface otherwise it filters and system prevents from sending it to the user interface. This unit uses a matching function and a decision to filter the output of matching function. There are two major approaches. (1) Filtering as Retrieval + thresholding. (2) Filtering as text classification [11].

2.3.1. Filtering as Retrieval + thresholding: A filtering system uses a retrieval algorithm to score each incoming document and delivers the document to the user if and only if the score is above a dissemination threshold. Some examples of retrieval models that have been applied to the adaptive filtering task are: Rocchio, language models, Okapi, and pseudo relevance feedback.

2.3.2. Filtering as text classification: A popular approach is to treat filtering as a text classification task by defining two classes: relevant vs. non-relevant. The filtering system learns a user profile as a classifier and delivers a document to the user if the classifier thinks it is relevant or the probability of relevance is high. The state of the art text classification algorithms, such as support vector machines (SVM), K nearest neighbors (KNN), neural networks, logistic regression and Winnow, have been used to solve this binary classification. Some machine learning approaches, such as logistic regression or neural networks, estimate the probability of relevance directly, which makes it easier to make the binary decision of whether to deliver a document.

The tasks of the filtering track in TREC included batch and routing filtering and adaptive filtering [12]. A batch filtering system uses a retrieval algorithm to score each incoming document. If the score is greater than a specified threshold, then the document is delivered to the user. The routing filtering systems are more similar to the retrieval systems, the profile remains constant and the task is to match an incoming stream of documents to a set of profiles. Both systems need to return a ranked list of documents. Adaptive filtering involves feedback to dynamically adapt IF systems [11]. The profile is adapted dynamically in the presence of feedback.

2.4. User Interface

After detecting accepted data, filtering unit send those to the user interface unit. User views him/her interesting information and can to read/save/send /print /edit / update data. Since user interesting is changing during time, system should receive new feedbacks from user and update user profile. There are two ways for user profile updating: (i) automatic forming though observation like browsing data, time that he / she is reading certain article or other types of information access patterns. (ii) Forming the profile through feedback that user expressing his/her preferences for an item, usually on a discrete numerical scale. So ratings may be gathered through explicit means, implicit means, or both [13]. Explicit ratings are those where a user is asked to provide an opinion on an item. Implicit ratings are those inferred from a user's actions. For example, evaluating the article the user has just read.

3. Associated Techniques In IF Systems Design

Information filtering system design usually involves one or more of a large number of technologies.

Commonly used technologies are: Traditional Information filtering techniques, Linguistic and learning filtering techniques, Data mining techniques, Clustering techniques, Graph base techniques.

3.1. Traditional IF (*IR techniques):*

Traditional IF most used of IR techniques for filtering because there are many similarities between IR and IF and They share many common tasks and techniques. The techniques that filtering use of them are: Boolean Model, Vector Space Model, and Probabilistic Model and Latent semantic indexing Model.

3.1.1. Boolean Model: This model is known as an exact-match model. In this model, Profiles are keywords that are constructed by using Boolean logic operations, such as, AND (conjunctive), OR (disjunctive) and NOT (negation).Profiles matches exactly with documents. If a document satisfies the Boolean expression, that document is deemed to be relevant; otherwise it is deemed irrelevant. The main advantage of this model is its simplicity, but it suffers particularly from strong limitations [14]: It is difficult (even impossible) to determine the difference between the most significant terms and those which are not, because all the words have the same weight and the same level of importance. Interesting documents may not be retained if they do not contain all the words describing a user's profile. In addition, a classification of retrieved documents by order of relevance is not possible.

3.1.2. Vector Space Models: The vector space model [15] is based on the statistical occurrence of terms in the profile (representing the user's information need) and

documents. The user profile consists of a collection of words, each with an associated weight, which occupies a slot in a vector. The incoming article is also viewed as a vector of weighted terms. The advantages of this approach are adaptability, robustness and minimal user intervention. The main disadvantages are the possibility of different terms in the article describing the same concept (synonymy) and the possibility of the same terms describing differing concepts based on differing context (polysemy), e.g. blind Venetian and Venetian blind. Thesauri have been used to overcome the problem of polysemy by expanding the initial query or profile. This has proved beneficial but has the disadvantage that the additional context provided by associated terms in a profile is ignored [16].

3.1.3. Probabilistic Model: The probabilistic model was first introduced by Maron and Kuhns in 1960 [17]. These models are based on probability ranking principle (PRP). The probability ranking principle [18] states that for optimal performance, documents in a collection should be ranked by the decreasing probability of their relevance to the request or information need, as calculated from whatever evidence is available in the system. Probabilistic IF model estimate the probability of relevance of documents for a profile.

3.1.4. Latent Semantic Indexing (LSI): Latent Semantic Indexing (LSI) [19] attempts to overcome the problems associated with word-base methods by organizing textual information into a semantic structure more suitable to information filtering. The LSI approach attempts to filter/receive information at a semantic, rather than at a syntactic or lexical level by not basing the comparisons between documents on the terms in document but on domains which these terms occur. This method suffers from other difficulties- such as problems in attaining fine-grained filtering without user-defined domains.

3.2. Linguistic and learning filtering techniques

This technique investigates the mechanisms by which knowledge is acquired through experience. In IF systems design, knowledge base techniques are commonly used to learn a model that is used to profile users as they use the system. The commonly linguistic and learning filtering techniques are: Rule-based IF, Neural Network base IF, Evolutionary genetic algorithms based IF, and case based reasoning IF.

3.2.1. Rule- based IF: Filtering systems can utilize rules to represent user profiles. Each rule can represent a user information need or pattern of information filtering. For example, in email messages, rules can be defined and applied to fields that appear in the message header (*e.g.*, sender, data sent, and subject).

Table 1. Traditional information filtering techniques

		Advantages	Disadvantages
Information retrieval techniques (traditional IF)	**Boolean Model**	• Simplicity	• It is difficult (even impossible) to determine the difference between the most significant terms and those which are not • Interesting documents may not be retained if they do not contain all the words describing a user's profile • A classification of retrieved documents by order of relevance is not possible
	Vector Space Model	• Adaptability • Robustness • Minimal user intervention	• It suffers from the semantic problems (Synonymy, homonymy, word ordering, etc.).
	Probabilistic Model	• Based on probability ranking principle (PRP) for optimal performance.	• In full text document databases, no experimental results are available for the applicability of probabilistic methods. • With multimedia documents, there is the problem of representation for non- textual parts. • The integration of text and fact retrieval will be a major issue. • PRP and many of methods for probabilistic model do scarcely take into account the special requirements of interactive information retrieval (IIR).
	Latent semantic indexing Model	• It can filter and select documents, which don't match any words with the user's interests. • It can be used to filter new information for more stable user's interest	• It has problems in attaining fine-grained filtering without user-defined domains. • The update operation of the concept space is expensive in time.

The rules may contain instructions on how to handle a message, depending on the values of these fields. For example: if the sender of an email message does not appear in a certain predefined list, the message gets a low relevance rank; if the subject of the message is about a certain topic, the message gets a high rank [2].

Table 2. Linguistic and learning filtering techniques

		Advantages	Disadvantages
Linguistic and learning filtering techniques	**Rule- based IF**	• Rule representation can be relatively easily applied for semi - structured data items.	• It is difficult to define rules that are significant to the user profile and whose parameters can be inferred at ease from unstructured data. • The gradual obsolescence of rules contained in the filters over time
	Semantic net IF	• Conceptual understanding of the text.	• Lower precision that could be addressed through the addition of index pattern.
	Neural Network base IF	• After the training phase, the network can be used as a black box to process new data. • This model is dynamic: it can learn and modify its behavior progressively.	• Main disadvantages of the NNs are their incapacity to explain the result which they provide.
	Evolutionary genetic algorithms based IF	• Evolutionary computation is applied to some problems cannot be solved by the traditional algorithms. They have been successfully applied to optimization and machine learning problems. • It allows the elimination of bad profiles and exploration of new domains which can interest user. • By combining evolutionary computation with other intelligent algorithms, the performance of the filtering technologies has been improved greatly.	• The basic weaknesses of classical evolutionary methods are high rate of iterations and significant computing expenses at their use. • Exploitation is a hard and difficult goal in EA.
	case based reasoning IF	• Knowledge acquisition process is considerably simplified. • Knowledge maintenance process is greatly facilitated.	• The similarity metric and adaptation is highly domain dependent • Matching process, if complex, can add computational cost to the CBR. • It can take a much larger number of cases than rules to cover a domain to the same extent.

3.2.2. Semantic-nets: Using semantic-nets techniques can alleviate some of rule base techniques difficulties. For example, an agent can use of semantic-net whose nodes are concepts and arcs are the co-occurrence relation of two concepts every node and every arc has a weight, reflecting users 'interests, dynamically updated in response to users, browsing activities. Some of these agents use semantic thesaurus repository.

3.2.3. Neural network based IF: A neural network is an inter-connected assembly of simple processing elements, units or nodes, whose functionality is roughly based on the animal neuron. The processing ability of the network is stored in inter-unit connection weights, obtained by a process of adaptation to, or learning from, a set of training patterns. The weights are supposed to adapt when the net is shown examples from training sets. Neural networks can also be applied in IF systems, where a user profile is representing a user's concept with unseen associations, that adapts from training [2].

3.2.4. Evolutionary genetic algorithms based IF: Evolutionary genetic algorithm based techniques borrow their model from the Darwinian concept of the natural process of survival. Nature selects the fit individuals to survive, and genetic patterns are passed by the individuals down through generations. The changes take place by recombining the genetic codes of pairs of individuals. These features allow us to apply an evolutionary and genetic approach in IF systems. The analogy in information filtering makes use of the vector space model to represent documents. In this model, a gene would be represented as a term, an individual as a document in the vector space, and the community as a profile. An appropriate objective function is introduced as the survival process, to decide whether to update the profile [2].

3.2.5. Case based IF: A case-based reasoning (CBR) system uses the recall of examples (cases) as the fundamental problem-solving process. It comprises a number of knowledge containers: the case-base, the vocabulary used to describe cases, the similarity measure used to compare cases and the knowledge needed to transform recalled solutions [20].

3.2.6. Other techniques: Other algorithms that use in this area are: K Nearest Neighbors (KNN) Algorithms, Bayesian Classifiers, The naïve Bayesian classifier, Belief (Bayesian) networks.

3.3-Web Data Mining Techniques

Data mining, or knowledge discovery in databases (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored in large databases, data warehouses, and other massive information repositories [21]. Data mining can use for user profiling. Zhou et al use after filtering, a data mining process based on the pattern taxonomy model (PTM) on the residual data set to rationalize the data relevance on the reduced data set [22]. Web mining is the application of data mining techniques to extract knowledge from Web .Web mining techniques can be apply for the Customization i.e. Web personalization. [23] Proposed a multi-agent system based on three layers: a user layer containing users' profiles and a personalization module, an information layer and an intermediate layer. They perform an information filtering process that reorganizes Web documents. Web mining is concerned with data mining on the Web. Many Web data mining methods have been developed to underpin IF system. For example, Web usage mining provides an excellent way to learn about users' interest [24].

3.3.1. Web Content Mining: Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. In Collaborative Filtering, for item-to-item correlation, an attempt is made to classify items based on their content or type of product, and recommend similar items to a customer. Web content mining approaches are either agent-based or make use of the database approach. Agent based approaches include: the use of intelligent search agents that conduct searches based on user profiles; IF and categorization agents that retrieve, filter and categorize documents; and the use of personalized web agents that learn user preferences and return personalized information to them based on their individual preferences. The database approach attempts to organize semi-structured web data into a more structured collection, suitable for database-like querying [20].

3.3.2. Web Usage Mining: Web Usage Mining refers to discovering user access patterns from Web usage logs. It focuses on using data mining techniques to analyze search logs to find interesting patterns [25]. The primitive object of Web usage mining is the discovery of Web access patterns. With Web usage mining, the user log can be analyzed. Some patterns about user behavior can be obtained from usage logs and then these patterns can be turned into a user profile. The user profiles are then utilized to filter incoming articles for the individual.

3.3.3. Web Structure Mining: Web Structure Mining refers to inferring useful knowledge from the structure of hyperlinks (in-links and out-links). It studies the model underlying the link structures of the Web. One of the applications of web usage mining is web page ranking in Google, *etc*. [25].

3.4. Clustering techniques

Clustering can be defined as the process of organizing objects in a database into clusters or groups such that objects within the same cluster have a high degree of similarity, while objects belonging to different clusters have a high degree of dissimilarity [26].

From the practical standpoint, it is difficult to conduct a comprehensive study on all existing clustering methods. Two traditional and well-known clustering algorithms are SOM and k-means. Other clustering techniques, such as fuzzy c-means, expectation maximization, etc., can be investigated.

The learning process of Self-organizing maps is based on a competitive and unsupervised artificial neural network. It is a clustering algorithm that is used to map high-dimensional data into a two-dimensional representation space. SOM can be used to explore the groupings and relations within such data by projecting the data onto a two dimensional image that clearly indicates regions of similarity [27].

The k-means algorithm is one of the most frequently used and simplest clustering algorithms because of its ease of implementation, simplicity, and superior capability and efficiency in dealing with large amounts of data. The k-means algorithm is a nonparametric approach that aims to partition objects into k different clusters by minimizing the distances between objects and cluster centers.

The basic k-means algorithm is composed of the following steps [28, 29].

- Randomly select k data items as the centers of cluster.
- Assign each data item to the group that has the closest centroid.
- When all data items have been assigned, recalculate the positions of the k centroids.
- If there is no further change, end the clustering task; otherwise return to step 2.

3.5. Graph Base Techniques

Graph structures used to represent web pages comprise nodes representing web pages, and arcs representing the strength of association between the web pages.

One of the most commonly used graph structures for web pages stores the hyperlink topology of a website or collection of sites. In this representation, directed arcs are used to connect one web page to another if a hyperlink links the former to the latter. This representation could be useful in recommender systems if there is the need to know if a given page p is reachable from a set of pages S, in which case, the task would be to

determine if an edge in the graph exists between any of the nodes present in S and Node p.

Another use of graph structures is to represent the similarities between web pages, or user behaviors.

In such a representation, arcs between nodes are labeled with the similarities between web pages. This representation could be useful in determining which pages to cluster together, and so treat them similarly.

Yet another graph representation could be used to show the frequencies of user accesses from one node to another. In this representation, directed arcs between nodes are labeled with the access frequencies between them.

4. Conclusion

Internet is a huge source of information. Information filtering allows us to automatically filter out the unwanted content of the information. In this paper we defined the analytical architecture of the web information filtering system and then, suggest a systematic framework to classify web information filtering structures. We hope this proposed framework will lead to empirical and technical comparison of web information filtering structures and development of more efficient structures at future.

Acknowledgements

This work is supported by Alzahra university of Tehran, computer engineering department. The authors are grateful to anonymous referees of this paper for their constructive comments.

References

[1] N. J. Belkin and W. B. Croft, "Information filtering and information retrieval: two sides of the same coin?", Commune. ACM, vol. 35, no. 12, **(1992)**, pp. 29–38.
[2] U. Hanani, B. Shapira and P. Shoval, "Information Filtering: Overview of Issues, Research and Systems", **(2001)** Kluwer Academic Publishers, Printed in the Netherlands.
[3] M. S. Vallim and J. M. A. Coello, "An Agent for Web Information Dissemination Based on a Genetic Algorithm", **(2003)** IEEE.
[4] D. Elliott, "An empirical analysis of information filtering methods", MSc(R) thesis, **(2011).**
[5] G. Cabanac, M. Chevalier, C. Chrisment, C. Julien and P. Tchienehom, "Web Information Retrieval: Towards Social Information Search Assistants".
[6] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval", First edition, Addison Wesley, **(1999).**
[7] D. Godoy and A. Amandi, "Modeling user interests by conceptual clustering", Information Systems, vol. 31, no. 4, **(2006)**, pp. 247–265.
[8] A. V. Leouski and W. B. Croft, "An Evaluation of Techniques for Clustering Search Results", Technical Report IR-76, **(1996)**.
[9] B. Shapira, P. Shoval and U. Hanani, "Stereotypes in Information Filtering Systems", Information Processing & Management, vol. 33, no. 3, **(1997)**, pp. 273–287.
[10] Y. H. Cho, J. Kyeong and S. H. Kim, "A personalized recommender system based on web usage mining and decision tree induction", Expert System with Applications, vol. 23, no. 3, **(2002)**, pp. 329 –342.
[11] Y. Zhang, "Adaptive Information Filtering", University of California Santa Cruz, **(2001)** by CRC Press LLC.
[12] S. Robertson and D. A. Hull, "The trec9 filtering track final report", in TREC-9, **(2000)**.

[13] J. B. Schafer, D. Frankowski, J. Herlocker and S. Sen, "Collaborative Filtering Recommender Systems", Springer-Verlag Berlin Heidelberg, **(2007)**.
[14] O. Nouali and P. Blache, "Automatic Classification and Filtering of Electronic Information: Knowledge-Based Filtering Approach", The International Arab Journal of Information Technology, vol. 1, no. 1, **(2004)** January.
[15] T. W. Yan and H. Garcia-Molina, "Index Structures For Information Filtering Under The Vector Space Model", Department of Computer Science, Stanford University, Stanford, CA, **(1994)**.
[16] H. Sorensen, A. O' Riordan and C. O' Riordan, "Profiling with the INFOrmer Text Filtering Agent", Journal of Universal Computer Science, vol. 3, no. 8 **(1997)**, pp. 988-1006 submitted: 1/6/97, accepted: 10/8/97, appeared: 28/8/97 - Springer Pub. Co.
[17] M. E. Maron and J. L. Kuhns, "Probabilistic indexing and information retrieval", On relevance, J. ACM, vol. 7, no. 3, **(1960)** July, pp. 216-244
[18] S. E. Robertson, "The probability ranking principle in IR", **(1997)**, pp. 281-286.
[19] S. Deerwester, S. Dumais, T. Landauer, G. Furnas and R. Harshman, "Indexing by latent semantic analysis", Journal of the Society for Information Science, vol. 6, no. 41, **(1990)**, pp. 391-407.
[20] D. L. Nkweteyim, "A collaborative filtering approach to predict web pages of interest from navigation patterns of past users within an academic website", PhD, University of Pittsburgh, **(2005)**.
[21] J. Han and M. Kamber, "Data mining: Concepts and techniques", San Diego, CA, USA: Academic Press, **(2001)**.
[22] X. Zhou, Y. Li, P. Bruza and Y. Xu, "Integration of Information Filtering and Data Mining Process for Web Information Retrieval", Proceedings of the 12th Australasian Document Computing Symposium, Melbourne, Australia, **(2007)** December 10.
[23] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization", ACM Transactions on Internet Technology (TOIT), vol. 3, no. 1, **(2003)**, pp. 1-27.
[24] J. Srivastava, R. Cooley, M. Deshpande and P. -N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data", SIGKDD Explorations, vol. 1, no. 2, **(2000)**, pp. 12–23.
[25] M. P. S. Bhatia and A. K. Khalid, "Information retrieval and machine learning: Supporting technologies for web mining research and practice", Webology, vol. 5, no. 2, **(2008)** June**.**
[26] A. K. Jain, M. N. Murty and P. J. Flynn, "Data clustering: a review", ACM Computing Survey, vol. 31, no. 3, **(1999)**, pp. 264–323.
[27] T. Kohonen, "Self-Organizing Maps", 3rd ed., Springer, **(2000)**.
[28] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm", Applied Statistics, vol. 28, no. 1, **(1979)**, pp. 100–108.
[29] M. David, "An example inference task: clustering information theory, inference and learning algorithms", Cambridge University Press, **(2003)** (Chapter 20).
[30] Q. Li and B. M. Kim, "Clustering approach for hybrid recommender system", Proceedings of the International Conference on Web Intelligence, Halifax, Canada, **(2003)**, pp. 33–38.

Authors

Narges Sadat Khozooii

She received her B.S. in Software Engineering from University of Shariati, Tehran, Iran. Currently, she is pursuing M.S. in Software Engineering at Alzahra University, Tehran, Iran. Her research interests include information filtering, information retrieval and recommender systems.

Saman Haratizadeh

He is an Assistant Professor at Tehran University, His main research interest is machine learning and its applications, especially in data mining domain and automated decision support. He also follows a line of research in biological inspired computation and emotion modeling in AI.

Mohammad Reza Keyvanpour

He is an Assistant Professor at Alzahra University, Tehran, Iran. He received his B.S. in Software Engineering from Iran University of Science &Technology, Tehran, Iran. He received his M.S. and Ph.D. in Software Engineering from Tarbiat Modares University, Tehran, Iran. His research interests include Data mining.