Mostafa, J., & Jacob, E. (1996). Influence of Classification on Information Filtering. 7th ASIS SIG/CR Classification Research Workshop, 103-106. doi:10.7152/acro.v7i1.12692

Proceedings of the 7th ASIS SIG/CR Classification Research Workshop

Influence of Classification on Information Filtering

Javed Mostafa
[jm@indiana.edu]
School of Library and Information Science
Indiana University, Bloomington, Indiana

and

Elin K. Jacob
[ejacob@indiana.edu]
School of Library and Information Science
Indiana University, Bloomington, Indiana

1. Introduction

Vast amounts of information are available via the Internet and there are indications that the volume of scholarly information shared across the Internet is increasing [Okerson, 1995]. Tools such as WAIS, Lycos and Webcrawler have been developed to allow users to access information, but these tools require the user to be actively involved in locating that information. With the explosive growth in digital and non-digital information, however, it has become increasingly difficult to find the time necessary to search for information.

To meet this challenge, systems are currently being developed that filter incoming information before it is presented to the user [Fischer & Stevens, 1991; Kilander, 1995; Maes, 1994]. Information filtering (IF) systems are similar to information retrieval (IR) systems in that they aid in document selection. But, as Belkin and Croft [1992] point out, an IF system differs from an IR system in that it assumes a dynamic document base that frequently contains semi-structured or unstructured items and can increase or decrease as new documents are received or old documents are discarded. In addition, document selection in an IF system relies on a long-term user interest profile instead of a search query.

Because the "information space" of a networked user may change in an unpredictable fashion, the Internet environment is especially suited to the application of filtering systems. We have developed a filtering system named SIFTER and have demonstrated that SIFTER can successfully filter typical e-mail documents generated by Internet listservs and mailing lists. In Mukhopadhyay et al. [1996], we investigated the influence on the filtering process of various parameters associated with the user interest profile. In the current work, we are investigating the influence of document classification methods on information filtering.

2. Statement of the Problem

A filtering operation establishes relevance for each document by comparing the document to the user's interest profile. Mapping between documents and interest profiles, however, can be extremely computationally demanding if the process requires computations over a high dimensional document space (number of unique keywords appearing in documents). The document classification component plays a crucial role in filtering in that it aids in transforming the original high dimensional document space to a more compact space consisting of a limited

ISSN: 2324-9773

Mostafa, J., & Jacob, E. (1996). Influence of Classification on Information Filtering. 7th ASIS SIG/CR Classification Research Workshop, 103-106. doi:10.7152/acro.v7i1.12692

number of document clusters. In this approach, the computational demand is significantly reduced because documents are ranked indirectly based on their class membership and the interest profile is specified in terms of classes instead of keywords [Mostafa et al., 1995]. Because document classification plays such an important role in filtering, a set of evaluative criteria are necessary to aid in comparing and ultimately selecting the appropriate classification technique.

We are comparing three classification techniques that differ in the amount of human involvement required in the determination of classes:

-- In the fully dependent mode of classification, the system relies upon class labels that are assigned by the human classifier. Labels are assigned to each document from a standard classification scheme and the process of machine classification consists of matching a class label in the document to one of the class labels maintained by the system.

-- In the partially dependent mode of classification, a supervised machine learning algorithm is used to train a neural network in a back-propagation fashion [Pao, 1989]. During training, the system is presented with both a set of classification labels taken from a standard classification scheme and a set of corresponding, unlabeled documents. The output of this supervised learning process is a configured neural network that has learned to classify unlabeled documents on the basis of the classification scheme used during training.

-- In the independent mode of classification, an unsupervised machine learning algorithm is used to discover classes from a set of training documents without class labels. This approach, known as maxi-min distance algorithm [Tou & Gonzalez, 1974] with cosine distance similarity measure, produces a set of cluster centroids which is used for classification.

By switching the classification module used in SIFTER, each one of these three different classification modes can be implemented.

3. Methodology

The study will use a randomly selected sample of 7500 documents from the domain of biomedicine that consist of title and abstract drawn . The document collection will be divided into two sets: one set consisting of 6000 documents will be used for learning in the supervised and unsupervised modes; the other set, consisting of 1500 documents, will be used for experimental testing.

The fully dependent classification mode is comprised of fifteen classes and will be used as the control level. The partially dependent mode of classification utilizes these same fifteen classes in training of the neural net. The independent mode of classification will generate its own set of classes based upon the same training set of 6000 documents used to train the neural net. However, the set of classes generated in the independent mode may or may not replicate the original fifteen classes in the control level.

Analysis of the classification process will evaluate document assignment in both the partially dependent and the independent techniques. Analysis of the filtering process will concern filtering performance of all three techniques: fully dependent, partially dependent and independent.

Analysis of document assignment will assume that the partitioning of the document set in the control mode is accurate. For each class in the control mode, we will compare the set of member documents to the class assignment of those same

ISSN: 2324-9773

Mostafa, J., & Jacob, E. (1996). Influence of Classification on Information Filtering. 7th ASIS SIG/CR Classification Research Workshop, 103-106. doi:10.7152/acro.v7i1.12692

documents in the partially dependent and independent modes. Document assignment in the partially dependent mode is expected to exhibit less dispersion than assignment to classes in the independent mode, since the latter technique does not utilize the control set of classes. Measures of dispersion will be based upon the number of documents in the original document set that are assigned to the same and/or related classes in either the partially dependent or independent modes.

Analysis of filtering performance will involve comparison of document ranking across the three modes of classification. The SIFTER system presently performs filtering by ranking incoming documents based on a user's interest profile. The three classification techniques outlined above will be utilized to create three different versions of SIFTER. Using the same simulated user profile and the experimental test collection of 1500 documents, each version of SIFTER will assign rankings to incoming documents and these rankings will then be compared across versions. Document ranking will be analyzed using normalized precision and recall as proposed in Salton & McGill (1983).

4. Possible Findings

It can be reasonably assumed that, with increased human input, the accuracy of classification will improve and lead, in turn, to improved filtering performance. Human dependent methods, however, come at some cost. In the context of filtering where the environment is dynamic, the classification method must be sufficiently robust so that it can be easily maintained and modified. With increased dependence on human input, this robustness may be compromised. The main goal, then, in comparing classification techniques at three different levels of human input is to identify the trade-off points between classification accuracy and filtering performance. Thus the research involves two levels of analysis. At the first level, overall classification accuracy and filtering performance will be compared across systems. At the second level, the relationship between classification accuracy and filtering performance within each individual system will be analyzed.

References

1. Belkin, N., & Croft, B. 1992. Information filtering and information retrieval: Two sides of the same coin? Communications of the ACM, 35(12), 29-38.
2. Fischer, G., & Stevens, C. 1991. Information access in complex, poorly structured information spaces. In Proceedings of the ACM Special Interest Group on Human Computer Interaction Annual Conference, pp. 63-70, April, 1991.
3. Kilander, F. 1995. A brief comparison of news filtering software. HTTP://www.dsv.su.se/~fk
4. Maes, P. 1994. Agents that reduce work and information overload. Communications of the ACM, 37(7), pp. 31-40.
5. Mostafa, J., Mukhopadhyay, S., Palakal, M., Hudli, A., Lam, W., & Xue, L. 1995. A multilevel approach to intelligent information filtering. Submitted to ACM Transactions on Information Systems.
6. Mukhopadhyay, S., Mostafa, J., Palakal, M., Lam, W., Xue, L., & Hudli, A. 1996. An adaptive multilevel information filtering system. In Proceedings of the Fifth International Conference on User Modeling, Hawaii, pp. 21-28, Jan. 2-5, 1996.

ISSN: 2324-9773

Mostafa, J., & Jacob, E. (1996). Influence of Classification on Information Filtering. 7th ASIS SIG/CR Classification Research Workshop, 103-106. doi:10.7152/acro.v7i1.12692

7. Okerson, A. 1995. About the ARL 5th edition of Directory of Electronic Publication. gopher://arl.cni.org:70/00/scomm/edir/edir95/about
8. Pao, Y. -H. 1989. Adaptive Pattern Recognition and Neural Networks. New York: Addison.
9. Salton, G. & McGill, J. 1983. Introduction to Modern Information Retrieval. New York: McGraw-Hill.
10. Special Issue on Information Filtering. Communications of the ACM, 35(12).
11. Tou, T. & Gonzalez, C. 1974. Pattern Recognition Principles. New York: Addison.

literature on filtering:

www.ee.umd.edu/medlab/filter

www.ee.umd.edu/medlab/filter

ISSN: 2324-9773