A NEW WORD-INTERSECTION CLUSTERING METHOD FOR INFORMATION FILTERING

Jun Lai, Ben Soh
La Trobe University Bundoora, VIC 3083, Australia

Mao Lin Huang
University of Technology, Sydney NSW 2007 Australia

Keywords: clustering, information filtering, information retrieval, search engine, World Wide Web.

Abstract: As the use of the web grows globally and exponentially, it becomes increasingly harder for users to find the information they want. Therefore, there is a need for good information filtering mechanisms. This paper presents a new, efficient information filtering method using word clusters. Traditional filtering methods only consider the relevance values of document. As a result, these conventional methods fail to consider the efficiency of document retrieval, which is also crucial. Our algorithm using offline computation attempts to cluster similar documents based on words shared by documents to produce clusters, so that the efficiency of information filtering and retrieval can be improved.

1 INTRODUCTION

The amount of information in the world is increasing far more quickly than our ability to process it. All of us have known the feeling of being overwhelmed by the number of new books, journal articles, and conference proceedings coming out each year. Now it is time to create the technologies that can help us sift through all the available information to find what is the most valuable and relevant to us in a more efficient way.

Currently there are some promising information filtering technologies:

- Content-based filtering: It is also called cognitive filtering. This system searches for items similar to those the user prefers based on a comparison of content using text-learning methods. Only the content and properties of a document contribute to the filtering, and each user operates independently. This is a traditional approach. This approach has difficulty capturing different types of content and has problem of over-specialization. When the system recommends items scoring highly against a user's preferences, the user is restricted to seeing items similar to those already rated.
- Collaborative filtering: It is also called social filtering. Here, documents are recommended for a user based on the likes of other users with similar tastes. User profiles are used to compare with each other. Groups of similar profiles are identified and users belonging to one group will be presented the same set of documents. The major drawback of this method is if the number of users is small or a user whose taste is unusual would not get high quality recommendation.
- Rule-based filtering: It uses demographic or other kind of purposely collected data of users to build user profiles and then define a set of rules to tailor the content delivery based on the facts specified in the user profiles. However, the creation and maintenance of rules are generally manual, as the system gets complicated, there will be difficulties managing it without conflict of logics.

Summarily, all current filtering systems consider only the relevance and importance to the users in different ways. However, as the system gets complicated, the efficiency becomes crucial. The surveys show that about 85% of Internet users make

Lin Huang M., Lai J. and Soh B. (2004).
A NEW WORD-INTERSECTION CLUSTERING METHOD FOR INFORMATION FILTERING.
In *Proceedings of the First International Conference on E-Business and Telecommunication Networks*, pages 241-244
DOI: 10.5220/0001382402410244

use of search engines and search service to find specific information. Users are not satisfied with the performance of the current generation of search engines because of slow retrieval speed, communication delays and poor quality of retrieved results [1].

In this paper, we propose a new efficient method called word-intersection clustering which can cluster more than two documents based on words shared by documents. This method applies an algorithm to compute the correlation similarity score of documents. The documents with the similarity score above a given threshold will be clustered together. A definition of documents profile is derived, so that each document has a profile based on the classification of category and similarity score. Then the documents are clustered under different categories. The proposed algorithm's offline computation scales independently of the number of documents. If one document in a cluster is relevant, then the whole cluster is relevant which makes the information retrieval more efficient.

This paper is organized as follows: The next section discusses the structure of a document based on the words shared by various documents. In section 3, we discuss the proposed algorithm and technique to cluster documents and the final section concludes the paper.

2 RESTRUCTURING OPERATION

Existing clustering methods focus on clustering two documents [2]. There has been a lack of effort on clustering more than two documents.

We propose a new restructuring operation by using those keywords appearing in the documents. Each keyword has different weight, ranging from 0 to 1. The value of weight is decided by system designer based on the importance and relevance of the keywords in that category and the number of times that keyword appears in that document.

Figure 1 shows the idea of restructuring operation of documents. The documents in the same category are clustered in accordance with the words shared by documents after the restructuring operation. For example, the documents 1, 15, 18 and 22 are clustered, documents 2 and 3 are clustered, and so are documents 7, 8 and 10.

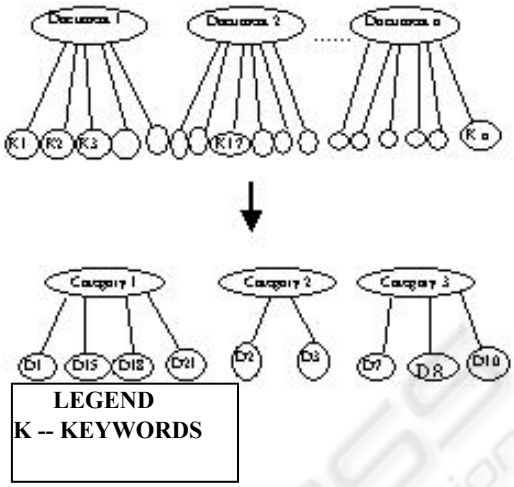

Figure 1: Restructuring operation of documents

3 DOCUMENT CLUSTERING

In this section, we discuss the algorithm and technique of word-intersection clustering.

We propose a restructuring operation to cluster documents as described in section 2. In this section, we will discuss the algorithm and technique of documents clustering. The subsections are organized as follows: section 3.1 presents the approach of calculation of similarity score. Then the document profiles will be derived in section 3.2. Finally, the proposed k-time clustering algorithm will be applied.

3.1 Computation of similarity score

To compute the similarity score of documents, first of all, we select some keywords appearing in those documents in a given category, whereby each word is assigned a weight, ranging form 0 to 1. Different word has different weight based on how important and relevant of that word is in a particular category. The value of weight is calibrated by system administration. For example in the category of information management, the words "information filtering" might be assigned by system designer to have higher weight than the words "data storage". The number of times a word appearing in a document also signifies the relevance value with respect to all other documents.

Table 1 shows the number of times a keyword appears in the document in the category of information management.

Table 1: Number of time that keywords appear in the documents

DID (document ID)	Keyword1	Keyword2	Keyword3	Keyword4
21	10	15	20	18
45	12	17	19	18
567	7	19	25	19

The similarity score is the sum of all product of keyword weight and the number of times that the keyword appears in the document. The similarity score computed by the following formula:

$$SS_{(d,c)} = \sum_{j=1}^{n} (W_{Kj} * Count_{Kj}) \quad (1)$$

where:

- SS is the similarity score of documents in a given category C based on keyword K.
- K j is the keyword in that document ($1 \leq j \leq n$).
- W_{Kj} is the pre-defined weight of the keyword K_j, determined by system admin.
- $Count_{Kj}$ is the number of times that keyword appearing in the document.

For instance, for the document 21 in the category of information management, the keyword 1 appears 10 times, while keyword 2 appears 15 times, keyword 3 appears 20 times, keyword 4 appears 18 times, Therefore, the similarity score of document 21 is:

$SS_{(21,info\ mgt)}$ = (0.8*10+0.7*15+0.5*20+0.6*18)

= 39.3

We can have table 2 based on formula (1).

Table 2: Similarity score of documents

DID	KW1	KW2	KW3	KW4	SS
21	10	15	20	18	39.3
45	12	17	19	18	41.8
567	7	19	25	19	42.8

3.2 Deriving document profile (DP)

From the calculation of similarity score, the document profile can be derived as follows:

$DP_{(d)}$ = {(c, $SS_{(d,c)}$ | c $\in C$, 0 ≤ $SS_{(d,c)}$ ≤ $SS_{threshold}$ } (2)

where:

- c denotes a category.
- C is all categories to which the document can be related.
- SS is the similarity score for document d.
- $SS_{threshold}$ is the minimum SS acceptable for a document to belong to that category.

From formula (2), each document can have a profile based on the classification of category and similarity score calculated by formula (1).

For example, the profile of document 21 is:

$DP_{(21)}$ = {(info mgt, 39.3), (knowlge mgt, 28), (data mining, 26), (data mgt, 13)}

3.3 Clustering Algorithm

Using the document profile, we can measure the correlation similarity score among documents. Table 3 shows the document profile.

Table 3: Document profile

DID	Info mgt	Knowlg mgt	Data mining	Data mgt
21	39.3	28	26	13
45	41	30	12	43
567	15	9	39	56

Table 3 also shows the similarity score of each document in different category. There are various clustering algorithms available; we chose K-mean [3]. We have defined our input data set for a general clustering already. Hence, any algorithm can be applied. K-mean algorithm splits a set of objects into a selected number of groups. The basic idea of K-mean is to find a single partition of the data, which has K number of clusters such that objects within the clusters are close to each other in some sense, and those in different clusters are distant. The object of clustering, in our case, is the document and the keyword appearing in the documents. Therefore, the documents in the same cluster will be considered as relevant to that category.

From the K-mean clustering, we will have K number of clusters. The documents belonging to the same cluster will have the relevant information. For example, if the given threshold is 25, then the document 21 is not relevant to the category of data management. The final pass of the algorithm produces the clustering of (21, 45) for category information management, (21, 567) for category data mining, (45, 567) for category data management.

4 CONCLUSION

In this paper, we propose a new word-intersection clustering method based on words shared by documents. This new method computes the correlation similarity score among documents. The document with similarity score above a given threshold will be clustered. Thereafter we derive the document profile based on the similarity score. Therefore, the document will be clustered for different categories. For the current information filtering methods, there has not been much focus on clustering more than two documents. Our approach computes similarity score and derives document profile offline. The documents have been pre-clustered which makes information retrieval more efficient. As future research, we would like to investigate if this method can be optimized.

REFERENCES

Kobayashi, M and Takeda, K, 1999. Information retrieval on the Web. In *ESSIR 2000*, LNCS 1980, Springer-Verlag, pp. 242-285.

Meng, X and Chen, Z, 2003. Personalized web search with clusters. In *IC'03, International Conference on Internet Computing*, pp. 46-52.

A Hartigan, J., 1975. *Clustering algorithms,* WILEY Publication.

Yang, F., Zhu, Y., Shi, B., 2003. A new algorithm for performing ratings-based collaborative filtering. In *Web Technologies and Applications: 5th Asia-Pacific Web Conference*, Springer-Verlag, pp. 239 – 250.

Breese, J., Heckerman, D., and Kadie, C., 1998. Empirical analysis of predictive algorithms for collaborative filtering. In *14th Conf. Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pp. 43-52.

Goldbeg. D., Nichols. D., Oki. B.M. and Terry. D., 1992. Using collaborative filtering to weae an information tapestry. *Communications of the ACM*, pp. 61-70

Goldberg K., et al., 2001. Eigentaste: a constant time collaborative filtering algorithm. *Information Retrieval Journal*, vol. 4, no. 2, pp. 133-151.

Mostafa, J., Mukhopadhyay, S., Palakal, M., and Lam, W., 1997. A multilevel approach to intelligent information filtering: model, system, and evaluation. *ACM Transactions on Information Systems*, Vol. 15, No. 4, pp. 368–399.

Smith, J., 1998. *The book*, The publishing company. London, 2nd edition.