

# Discovery of Aggregate Usage Profiles based on Clustering Information Needs

Azreen Azman  
Department of Computing Science  
University of Glasgow  
Glasgow, Scotland, G12 8RZ  
azreen@dcs.gla.ac.uk

Iadh Ounis  
Department of Computing Science  
University of Glasgow  
Glasgow, Scotland, G12 8RZ  
ounis@dcs.gla.ac.uk

## ABSTRACT

We present an alternative technique for discovering aggregate usage profiles from Web access logs. The technique is based on clustering information needs inferred from users' browsing paths. Browsing paths are extracted from users' access logs. Information need is inferred from each browsing path by using the Ostensive Model[1]. The technique is evaluated in a document recommendation application. We compare the performance of our technique against the well-established transaction-based technique proposed in [2]. Based on an initial evaluation, the results are encouraging.

**Categories and Subject Descriptors:** H.3.3 [Information Search and Retrieval]

**General Terms:** Algorithms, Performance

**Keywords:** Ostensive Model, Clustering Information Needs, Recommendation, Aggregate Usage Profiles

## 1. INTRODUCTION

A Web document recommendation system is intended to help a user in browsing by suggesting interesting or relevant documents interactively. The aim of such a system is to recommend a set of documents on the basis of previously visited documents.

In the first stage of recommendation, a set of aggregate usage profiles is generated where each profile contains a set of documents with similar access patterns. Secondly, a set of unvisited documents is recommended based on documents already visited by a user.

The performance of such a recommendation system relies on the quality of the generated aggregate usage profiles. Most proposed techniques in the literature use only access information in generating such profiles[2][3]. In this paper, we propose a technique that uses both content and access information for aggregate usage profile generation. We assume that by combining content and access information, a better set of aggregate usage profiles will be generated and the recommendation performance will be improved.

## 2. AGGREGATE USAGE PROFILES

User sessions are extracted from web access logs, where each user session consists of all documents visited by a user

in one session. A set of browsing paths,  $T$ , is extracted from the collection of user sessions, where each browsing path,  $t_v \in T$ , consists of all documents visited sequentially by following hyperlinks.

We present two techniques for discovering aggregate usage profiles, namely the *transaction clustering* and the *information need clustering* techniques, respectively. Transaction clustering is used as a baseline and is adopted from [2].

### 2.1 Transaction Clustering

Each browsing path,  $t_v \in T$ , is represented as an  $n$ -dimensional vector where each dimension is assigned with a document's weight,  $\vec{t}_v = \{(d_i, w_{iv})\}$ ,  $i = 1, 2, \dots, n$ , where  $n$  is the total number of documents;  $w_{iv} = 1$  if  $d_i \in t_v$  and  $w_{iv} = 0$  otherwise. The collection of browsing paths,  $T$ , is then clustered into a given number of clusters.

Each cluster  $c_l$  consists of a set of browsing paths,  $T_l$ , where  $T_l \subseteq T$ . The probability of occurrence of each document in each cluster is given by  $P(d_i, c_l) = \frac{|T_{li}|}{|T_l|}$  where  $T_{li} \subseteq T_l$ ,  $t_i \in T_{li}$  and  $d_i \in t_i$ . A document,  $d_i$ , is assigned to a cluster,  $c_l$ , if  $P(d_i, c_l) \geq \phi$ , where  $\phi$  is a threshold value. The value of  $\phi$  is set experimentally.

### 2.2 Information Need Clustering

A collection of information needs,  $N$ , is extracted from a set of browsing paths,  $T$ . An information need,  $n_v \in N$ , is inferred from each browsing path,  $t_v \in T$ . Information need extraction is based on the Ostensive Model[1]. In the model, an information need,  $n_v$ , is a set of features  $f_k$  and their weights,  $w_k$ ,  $n_v = \{(f_k, w_k)\}$ ,  $k = 1, 2, \dots, m$ , where  $m$  is the total number of features indexed. A feature is a keyword in a Web collection.

Documents are associated with weights,  $ORel$ , reflecting their age, i.e. the position in a browsing path. Three 'aging' assumptions are proposed; As the age of a document increases, its degree of importance *increases*, is *equal* or *decreases*. These assumptions are modelled by using functions,  $2^j$ ,  $\frac{1}{l_p}$  and  $\frac{1}{2^j}$ , respectively, where  $j$  is the position of document in  $t_v$  and  $l_p = |t_v|$ . The overall weight of a feature,  $f_k$ , in an information need,  $n_v$ , is calculated as the weighted sum of the  $ORel$  weights and feature weights over all documents in  $t_v$ ,  $w_k = \sum_{j=1}^{l_p} ORel_j \times w_{kj}$ , where  $w_{kj}$  is a weight of feature,  $f_k$ , in the document at position  $j$ .

Given that a function  $2^j$  is used, the weight of the document at position  $j$  in a browsing path, is given by  $ORel_j = 2^j$ .  $ORel_j$  is normalized such that  $\sum_{j=1}^{l_p} ORel_j = 1$ .

$N$  is clustered into a given number of clusters. Given that

$cen_l$  is the centroid of cluster  $c_l$ . Each document,  $d_i$ , is assigned to a cluster,  $c_l$ , if  $Sim(d_i, cen_l) \geq \alpha$ , where  $\alpha$  is a threshold value. Cosine similarity of two vectors could be used as  $Sim$  measure. The value of  $\alpha$  is set experimentally.

### 3. RECOMMENDATION AND EVALUATION MODEL

Given a browsing path of a user,  $t_e$ ,  $s$  number of documents at the beginning of  $t_e$  are selected as an active session set,  $S$ , where  $S \subseteq t_e$  and  $|S| = s$ . We assume  $S$  to be a set of visited documents. The model will recommend a new set of documents,  $RS$ , where  $RS = \{d_i\}, 1 < i < n$ , if  $Rscore(S, d_i, c_l) \geq \beta$ ,  $\beta$  is a recommendation threshold and  $n$  is the total number of documents. The recommendation score,  $Rscore(S, d_i, c_l)$ , is given by:

$$Rscore(S, d_i, c_l) = \sqrt{match(S, c_l) \times weight(d_i, c_l)} \quad (1)$$

where  $weight(d_i, c_l)$  is the weight of a document  $d_i$  in a cluster  $c_l$  and  $match(S, c_l) = Sim(S, c_l)$ . In the information need clustering technique, an information need  $n_S$  is inferred from  $S$ . Thus,  $match(S, c_l) = Sim(n_S, cen_l)$ . If a document appears in more than one cluster, the maximum  $Rscore$  of the document is chosen.

The performance of a recommendation set,  $RS$ , for a given browsing path,  $t_e$ , is evaluated based on *precision*, *coverage* as well as  $F1$  and  $R$  measures. Precision and coverage measures are given by  $precision(RS, t_e) = \frac{|RS \cap (t_e - S)|}{|RS|}$  and  $coverage(RS, t_e) = \frac{|RS \cap (t_e - S)|}{|(t_e - S)|}$ , respectively.  $F1$  and  $R$  measures are computed as follows:

$$F1(RS, t_e) = \frac{2 \times precision(RS, t_e) \times coverage(RS, t_e)}{precision(RS, t_e) + coverage(RS, t_e)} \quad (2)$$

$$R(RS, t_e) = \frac{coverage(RS, t_e)}{|RS|} \quad (3)$$

### 4. RESULTS

The evaluation is conducted by using the *Music Machines* websites collection[3]. The collection consists of a set of documents and access logs from March 1997 to April 1999. Six months data (March 1997 to August 1997) are used for training to generate aggregate usage profiles. One month data (September 1997) is used for evaluation. In both techniques, the *K-means* clustering algorithm is used.

We use  $F1$  and  $R$  measures. We would expect higher value of  $F1$  and  $R$  to indicate a better recommendation performance.  $F1$  is a single value measure for precision and coverage as traditionally used in Information Retrieval evaluation. For the  $R$  measure, coverage is normalized by the size of a recommendation set.  $R$  measure is the best performance measure for this particular recommendation problem, since we are interested in good performance with a smaller size of recommendation set.

In the experiment, **S1** (*Decreasing Weight*), **S2** (*Equal Weight*) and **S3** (*Increasing Weight*) are based on the information need clustering technique where  $Orel$  is  $2^j$ ,  $\frac{1}{j}$  and  $\frac{1}{2^j}$  respectively (See Section 2.2). For instance, recently visited document is weighted less in **S1** and is weighted more in **S3**. **S4** (*Transaction Baseline*) represents the transaction

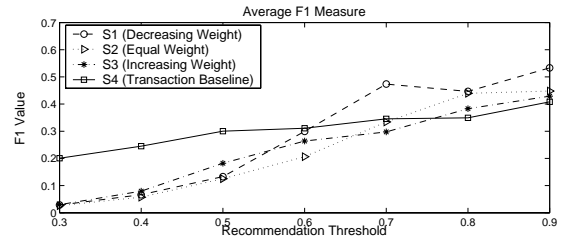


Figure 1: Average F1 measure

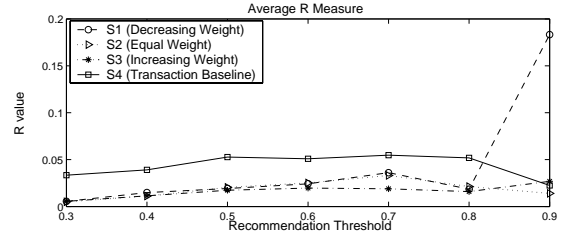


Figure 2: Average R measure

clustering technique (See Section 2.1). Figure 1 shows that our techniques (**S1**, **S2** and **S3**) perform better than the baseline (**S4**) at higher recommendation thresholds (threshold  $\geq 0.8$ ) for the  $F1$  measure. However, the baseline technique generally performs better for the  $R$  measure (Figure 2) at all recommendation thresholds except 0.9. The dramatic improvement in the  $R$  measure of **S1** at threshold 0.9 is due to a very small number of browsing paths being recommended. Our techniques tend to produce a larger size of recommendation set. This may be due to a higher weight assigned to each document in the clusters produced by **S1**, **S2** and **S3** as compared to **S4**.

### 5. CONCLUSION

In conclusion, the results are nevertheless encouraging, showing that our technique is comparable to the traditional document recommendation model proposed in [2]. However, there is still room for further improvement. Indeed, based on the results, we have identified a number of avenues for further investigation.

In the future, a refined technique that includes normalization of document weights in the cluster, may reduce the size of the recommendation set. Also, better clustering techniques such as Support Vector Machines (SVM) could be employed to produce better clusters. Finally, better and bigger collections could be used for the experiments to check for collection-independency.

### 6. REFERENCES

- [1] I. Campbell. *The ostensive model of developing information needs*. PhD thesis, University of Glasgow, September 2000.
- [2] B. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Discovery and evaluation of aggregate usage profiles for web personalization. *Data Mining and Knowledge Discovery*, 6:61–82, 2002.
- [3] M. Perkowitz and O. Etzioni. Towards adaptive web sites: Conceptual framework and case study. *Artificial Intelligence*, 118(1-2):245–275, April 2000.