EUROGRAPHICS Workshop on Multimedia (2004)
N. Correia, J. Jorge, T. Chambel, Z. Pan (Editors)

A New Approach to Filtering Multimedia Information Based on its Structure

Xiaodi Huang[1] and Jianming Yong[2]

[1]Department of Mathematics and Computer Science, Faculty of Science
[2]Department of Information Systems, Faculty of Business
University of Southern Queensland, Toowoomba, QLD, 4350, Australia

Abstract

In information filtering systems, the multimedia documents are sequentially presented to users based on the user relevance values. This paper argues that the presented multimedia documents should be both important and relevant to the users. The importance of a document is determined by its relations to others in the collection. All users are supposed to look for important and relevant documents. Based on this view, a structure-based filtering framework is described, which incorporates the characteristics of the importance and relevance of multimedia documents. An approach to calculating importance values of multimedia documents and then combining them into relevance values of multimedia documents is proposed to improve the representation of user profiles. An example is provided.

Categories and Subject Descriptors I.3.3 [Computer Graphics]: Display algorithms, H.3.1 [Information Storage and Retrieval]: Information Search and Retrieval-Information Filtering.

1. Introduction

As the World Wide Web grows exponentially, it becomes more and more difficult for users to find the information they want. In order to reduce this information overload, it is useful to prioritize the information. Such prioritizing can take the form of highlighting highly important items or deleting ones that are not considered relevant. Information filtering is concerned with such an information identifying process in which documents are selected from a stream of incoming data to satisfy a relatively stable and specific information need. It traditionally falls into three categories [MAL87]:

- Content-based filtering (also called cognitive filtering): where documents are selected on the basis of the correlation between the content of documents and users' preferences. Only the content and properties of a document contribute to the filtering, and each user operates independently. This is a traditional approach.

- Social filtering (also called collaborative filtering): where documents are filtered for a particular user based upon the preference of other users with similar tastes. User profiles are used to compare against each other. Groups of similar profiles are identified and users belonging to one group will be presented the same set of documents. Social filtering systems need a number of participants and documents to efficiently work together. This is the major drawback.

- Economic filtering: where documents are filtered by considering cost factors. Such factors can be the relation between cost and benefit of use, or the available network bandwidth and size of documents.

Hanani, Oard and Belkin et al. [HSS01] [Oar97] [BC92] believe that a "good" information filtering can successfully indicate the relevance of incoming documents, and thus protect users from irrelevant information, and without missing relevant information.

In summary, the above three kinds of filtering systems consider only the relevance to users or cost of documents in different ways. However, a document maybe be relevant, but not important, or vice versa, to the users. All the users are reasonably assumed to be presented both important and relevant documents. In general, a document in the collection has the following characteristics:

© The Eurographics Association 2004.

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

- Importance

Importance indicates that the degree of the role of a document plays in a whole document collection. Different documents have unequal roles. Some are influential while the others are trivial. Users may access to a set of documents which they consider to be relevant. They wish to automatically rank these documents in terms of "importance", and then to deal with those important ones in a priority manner. For example, we maybe survey scientific literature, looking for papers on information retrieval. Of course, we want to read the most influential papers firstly. In other words, we are concerned not only with the relevant content, but also with their important impacts in a large volume of relevant information.

- Relevance

Relevance is a confusing and much debatable concept. The generally accepted theoretical conceptualization of relevance involves the relationship between a user's information problem or need and the information that can solve the problem. The operational conceptualization involves a user's decision to accept or reject retrieved information from an information system [Lin94].

Indeed, the relevance of a document is related to its importance in some cases. The relevance, however, is traditionally treated as a concept relating to users. It is represented as an integral part of users' profiles. Here, we regard importance as the inherent characteristic of documents and a common component of all users. In other words, every user intends to access important information. The importance involves the relationships between the documents in a collection, while the relevance indicates the relationships between the documents and users.

Based on the above view, we argue that a filtering system should provide users with relatively important as well as highly relevant documents. In this paper, we propose a new approach called a structure-based filtering, which combines both the importance and relevance values of a document, particularly for a multimedia document containing such elements as images, video clips, and audios. Links between multimedia documents as well as between their multimedia elements provide a natural mechanism for quantifying the notion of "importance". More specially, a link can indicate the judgment of the author of one multimedia document as to the importance of another multimedia document. The proposed approach initially extracts the link structure of a multimedia document collection. The importance values of multimedia documents are then quantified by extending a notion of "centrality" widely used in social network analysis. Finally, all multimedia documents in the collection are ranked by their overall ranking scores, which are calculated with incorporation of both the importance and relevance aspects of the multimedia documents.

The rest of this paper is organized as follows. A structure-based filtering framework is presented in the following section, followed by describing a way of ranking nodes in a graph in Section 3. The multi-relation graph modeling is described in Section 4. Section 5 considers a combination of ranking documents with incorporating their importance. Following this, an algorithm is provided in Section 6. An example is shown in Section 7. After reviewing related work in Section 8, the conclusion is drawn in Sections 9.

2. A Structure-based Filtering Framework

In this section, we present a structure-based filtering framework based on the view discussed in the previous section.

The overall problem of information filtering can be broadly described as learning a map from a space of multimedia documents to a space of user relevance values. More precisely, we denote the space of multimedia documents with a number of k attributes as D and the space of user relevance values as R. The objective is to learn a map $f : D \to R$ such that $f(d)$ corresponds to the ranking score of a multimedia document d. Given that such a map is known for all points in D, a finite set of multimedia documents can always be rank-ordered and presented in a prioritized fashion to the user [MMP*97].

In our framework, the map is decomposed into two levels at two parallel parts as shown in Figure 1. The higher level in the first line represents a structure mapping f_{11} from a multimedia document space to a number of connected, weight graphs. With these graphs, the nodes represent the multimedia documents, and the edges represent relationships between the multimedia documents with respect to the k kinds of relations. That is, $f_{11} : D \times D \to G_k(V_k, E_k)$, where k is an integer denoting the number of possible relations among documents in a collection. This mapping is learned in an off-line setting, based on the link analysis of the collection of multimedia documents. The lower level in the first line subsequently employs another mapping f_{12} from the structure graph G_k to a set of importance values of nodes in the graph G_k, which measure how important roles multimedia documents play in the collection, i.e., $f_{12} : V_k \to R_n$.

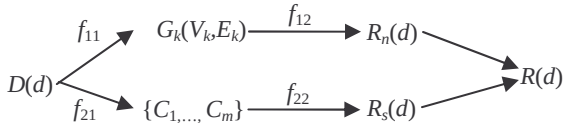


Figure 1: The mappings of the structure-based filtering

© The Eurographics Association 2004.

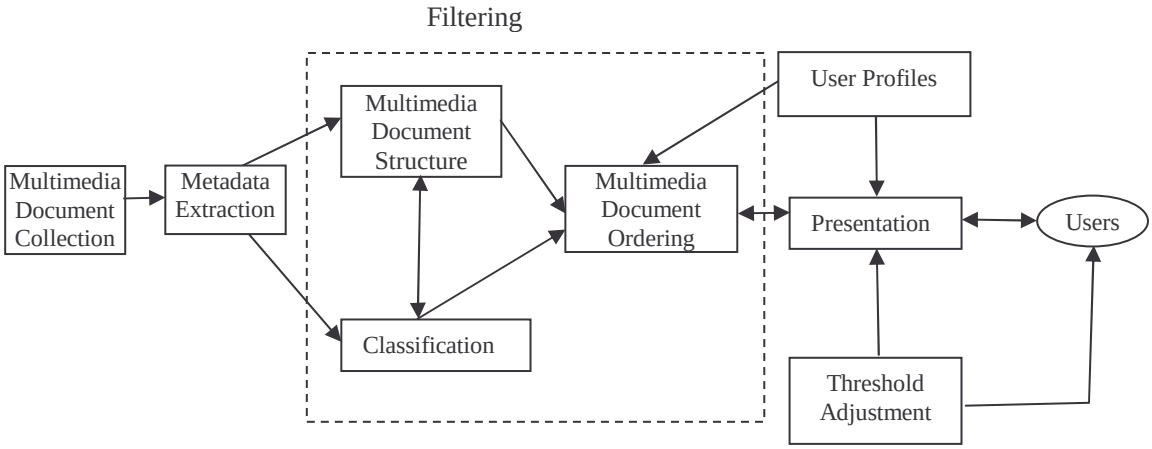


Figure 2: The Conceptual architecture of structure-based filtering systems

In the second line, the higher level partitions the multimedia document space into *m* classes by using a clustering technique, i.e., $f_{21}: D \rightarrow \{C_1, \cdots, C_m\}$. The lower level then estimates the mapping f_{22} describing user relevance for the different classes, i.e., $f_{22}: \{C_1, \cdots, C_m\} \rightarrow R_s$ [MMP*97]. The whole mapping *f* will be integrated by a combination of all those mappings. The value $R_n(d)$ indicates the importance of a particular *d* document, which has nothing with the characteristics of a particular user. In other words, the importance of a document is determined by its relations to others in the collection. The $R_s(d)$ is the user relevance value, and $R(d)$ denotes the ranking score of document *d* in the collection.

Figure 2 shows a conceptual architecture of structure-based filtering systems. Basically, it is composed of four components: *Multimedia Document Collection*, *Metadata Extraction*, *Filtering Engine* and *User Profiles*.

Multimedia Document Collection: This might be web sites, a set of databases, and email folders.

Metadata Extraction: This extracts potentially relevant information, and passes it to a filter engine. A relation between multimedia documents can be established on a basis of various attributes of multimedia documents, such as whether there exists a reference between one multimedia document and another or they have common keywords. With this information, the characteristics of the multimedia document collection can be derived and analyzed.

Multimedia Document Structure: Based on the information from *Metadata Extraction*, *Multimedia Document Collection* can be represented as a graph, where a node represents a multimedia document and an edge represents the link relationship between them.

Classification: The multimedia documents are classified into a finite number of classes by using a clustering technique.

Multimedia document Ordering: The multimedia documents are ranked by their ranking scores, which combine both importance and relevance values of multimedia documents.

User Profiles: A vector is used to represent user's preference. More importantly, user profiles are used for comparing multimedia documents to find those important and relevant ones, and also for grouping the users. The importance values of multimedia documents are combined into user files as a common feature of users.

Presentation: The multimedia documents are sequentially presented to users in a priority form.

There is also a feedback mechanism in the framework to improve the performance of the system.

There are already many techniques for clustering multimedia documents in the literature. We will present an approach to implementing the upper line mappings shown in Figure 1.

The above framework is based on an underlying assumption that all users need the importance and relevance information. It is possible that documents are relevant to the users, but not significant, or documents with equivalent relevance to the users are not equally important. The importance of a document is concerned here with its relations to others in a collection. In other words, it is a relation between documents in a particular collection. In contrast, the relevance of a document is about a relation between documents and users, thereby depending on the characteristics of both documents and users. In some cases, a document can be highly related to a particular user, while other users are not interested in it at all. However, an important document in a collection is always prominent, entirely independent of user profiles. This fact leads to a question of how to measure the importance of a document in a given

© The Eurographics Association 2004.

collection. As we know, a document can be represented a vector of its attributes. Ranking documents in a collection is equivalent to ranking their attributes. Different attributes reflect various aspects of a document. One document may be important in terms of one attribute, but unimportance with respect to other attributes. We should therefore compare these attributes with each other separately. A further question of quantifying the importance of document attributes arises. In the following, we present an approach to solving this problem where documents are modelled into a number of weight graphs associated with their different attributes. We begin with presenting a way of ranking nodes in a graph.

3. *NodeRank*: Ranking Nodes in a Graph

As mentioned before, the relationships among multimedia documents in a collection can be illustrated as a connected graph regarding their each kind attribute. With this graph, we need a function, namely f_{12}, to map every node of the graph into a positive real value as its importance. This value indicates the important position of a corresponding multimedia document in the collection with respect to this particular attribute. Fortunately, we can employ the measure of "centrality" used in social network analysis as a basis of the mapping.

3.1 Centrality of a Node in a Graph

Centrality refers to the importance of a particular node in a network. The measures of centrality have been developed to "attempt to describe and measure properties of 'actor location' in a social network" [WSK94]. In a multimedia document collection, the relationships between multimedia documents are represented as a graph. The importance of a multimedia document in such a graph can be measured by multimedia documents passed through it, or it can easily reach other multimedia documents in the collection. Or it is itself directly connected to other multimedia documents. From this perspective, the role of a multimedia document can be a function of its position in a given collection.

There currently exists a variety of centrality measures. These measures are roughly classified into three major types: degree, closeness, and betweenness centrality, which are respectively defined as follows.

Definition 1 (Freeman) *Degree Centrality*: The number of edges attached to a node *u*.

Obviously, Degree centrality can be normalized to range from 0 to 1, where 0 means the smallest possible centrality and value 1 the highest possible centrality. The normalized measures are called relative measures of centrality:

$$C_1(u) = \frac{C_1'(u)}{n-1}$$

where $C_1'(u) = |\{v \mid (u,v) \in E \wedge v \in V\}|$, and $n=|V|$.

Degree centrality reflects the direct relationships of a multimedia document with other multimedia documents in the collection.

It indicates the number of links connecting adjacent nodes to a local node, so it is a local centrality.

Definition 2 (Freeman) *Closeness Centrality*: The sum of geodesic distances, defined as the shortest path connecting two nodes, between a node *u* and all other nodes.

$$C_2(u) = \frac{\sum_{v \in V} d(u,v)}{n-1}$$

where $d(u, v)$ is the shortest path between nodes u and v, which is equal to the number of edges between them. Measures of centrality based on closeness reflect a node's freedom from the controlling actions of others, their capacity for independent action within the network. This measure actually indicates how far a node is from all others. A node with a higher closeness score is less centralized one than a node with a lower closeness score. The most central nodes can quickly interact with all other nodes because they are close to all others.

Definition 3 (Freeman) *Betweenness Centrality*: The ratio of the number of shortest paths between two nodes passing a node *u* to the number of all possible such shortest paths in a graph:

$$C_3'(u) = \sum_{j=1}^{n} \sum_{k=1}^{j-1} \frac{g_{jk}(u)}{g_{jk}}, \quad C_3(u) = \frac{C_3'(u)}{(n-1)(n-2)/2}$$

where g_{jk} is the number of shortest paths from node *j* to node *k*, and $g_{jk}(u)$ is the number of those shortest paths that include node *u*.

Centrality measures based on betweenness reflect the intermediary location of a node along indirect relationships linking other nodes. Betweenness centrality "measures the extent to which a particular point lies 'between' the various other points in a graph: a point of relatively low degree may play an important 'intermediary' role and so be very central to the network" [SJ00]. A node with high betweenness has a capacity to facilitate or limit interaction between the nodes it links.

Closeness and Betweenness are the global centralities.

3.2 Ranking a Node in a Weight Graph

The above centralities have at least two limitations. First, they are not suitable for a weight graph. Second, different above centrality measures may give quite different results for the same graph. It can be a case in which a node has a low degree, but with a high betweenness centrality. To remedy the first problem, we extend the three centralities to

© The Eurographics Association 2004.

cater for a weight graph. Suppose a given graph with weights denoted by G= (V, E, W), where every edge is attached with a value within [0, 1] indicating the degree of two linked nodes, we have ($E \to \Re$):

Degree Centrality

$$C_1(u) = \frac{C_1'(u)}{n-1} = \frac{\sum_{(u,v)\in E} w_{uv}}{n-1}$$

Closeness Centrality

$$C_2(u) = \frac{\sum_{v\in V} d(u,v)}{n-1} = \frac{\sum_{v\in V} \min\{\sum w_{uv}\}}{n-1}$$

where w_{uv} is the weight of an edge connected nodes u and v.

As defined previously, the *Betweenness Centrality* is concerned with the number of the shortest paths between two nodes, which is independent of the weights of edges. This leads to that it remains unchanged in the context of weight graphs.

In order to overcome the drawback of each single measure, a linear combination of degree, closeness and betweenness yields the following measure:

$$R_n(u) = \sum_{i=1}^{3} \alpha_i C_i(u) \quad (1)$$

where the weights α_1, α_2 and α_3 sum to 1. For simplicity, we can give equal importance to the three measures by assigning equal weights.

We take a weight graph shown in Figure 3 as an example, where the three centralities of node N_3 are computed as follows:

Degree: $C_1(N_3) = (0.4 + 0.3 + 0.75)/(7-1) = 0.21$

Closeness:

$$C_2(N_3) = \sum_{i=1}^{7} d(N_3, N_i)/(7-1)$$
$$= [0.4 + 0.3 + 0 + (0.75 + 0.6) + 0.75 + (0.75 + 0.2) + (0.75 + 0.25)]/6 = 0.792$$

Beteweeness:

$$C_3'(N_3) = \sum_{j=1}^{7} \sum_{k=1}^{j-1} g_{jk}(N_3)/g_{jk}$$
$$= (2+3+3+3+3)/(1+2+3+4+5+6) = 14/21$$
$$C_3(N_3) = 2C_3'(N_3)/(7-1)(7-2) = 0.044$$

Given $\alpha_1 = \alpha_2 = \alpha_3$ and $\alpha_1 + \alpha_2 + \alpha_3 = 1$, the *NodeRank* of node N_3 is

$$R_n(N_3) = \sum_{i=1}^{3} \alpha_i C_i(N_3) = 0.349$$

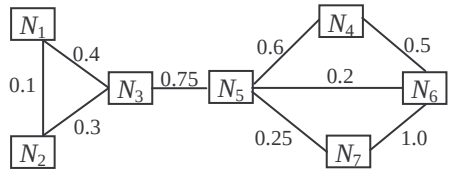


Figure 3: An example of the calculation of centrality index

Based on the above description, we give the following definition 4.

Definition 4 *NodeRank*: Let $G = (V, E)$ be an undirected, weighted, and connected graph. Let f_{12} be a function which assigns a positive real value to a node of G. The *NodeRank* of node u is denoted by $R_n(u)$, and $0 \le R_n(u) \le 1$.

In general, a node has a high *NodeRank* score, if it has a high degree, is easily accessible to (close to) all other nodes, and lies on several geodesics (the shortest paths) between other nodes.

4. The Multi-relation Graph Modeling of a Multimedia Document Collection

We model a multimedia collection as a graph G, where nodes represent documents, and edges indicate relations between two documents. The *NodeRank* of a document is based on its relation to others in a graph. This means ranking all the documents relies only on this kind of relation. In fact, a document has many attributes such as *Hyperlink*, *Reference*, and *Keyword*. A document can be represented by these attributes. The importance of a document in a collection can therefore be obtained by comparing its attributes with those of other documents. This means ranking documents becomes ranking their attributes. Two documents can have a relation with respect to each common attribute. In other words, we can model a collection into multi-graphs corresponding to different attributes, as illustrated in Figure 4. Therefore, the *NodeRank* of a document is determined by comparing a set of attributes with those of other documents, rather than by a single attribute as one graph.

Note that an attribute may have sub-attributes. For example, if Web documents are thought of as having three attributes, *Hyperlink*, *Reference* and *Keyword*, the latter is actually composed of a set of specific keywords, called a sub-attribute vector. A combination of sub-attributes is used to build the relations and the weights of two nodes in its corresponding attribute graph.

Apart from the multi-relations with other documents, the importance of a document also depends on the degree of these relations. Equally, the document importance should be measured by multi-relation weight graphs. This raises a

© The Eurographics Association 2004.

question of how to calculate the weights.

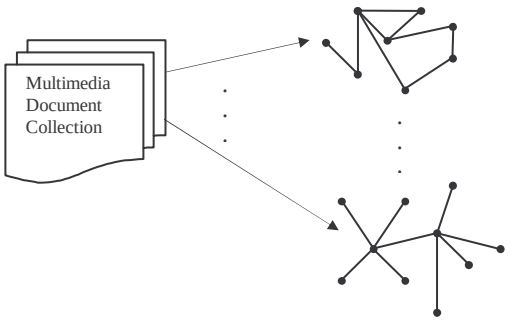


Figure 4: The multi-relation weight graph model

Here we present two ways of obtaining the relationship weights between two documents. One way simply counts the frequency of relations occurring in two corresponding documents, and then normalizes it. Formally, we have

$$w_{uv}^{k} = \frac{\#(a_{uv}^{k})}{\max_{(u,v)\in D}\{\#(a_{uv}^{k})\}}$$

In other words, the weight in *k*-attribute graph is the number of an attribute *k* occurring in both documents *u* and *v*, divided by the maximum number occurring in two documents in a collection.

In the case of Figure 5 where two multimedia documents with two linked elements, Text and Image, we specify 0.5 in a "link" attribute graph as the weight of the edge linking Document 1 and Document 2, if the maximum number of links in the collection is 4.

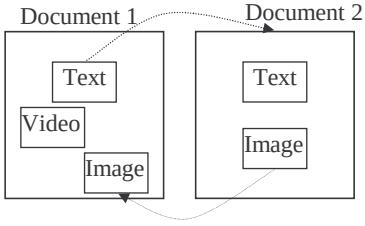


Figure 5: Construction of structure links in multimedia documents

The second approach uses the well-known vector model. For example, in a *Keyword* attribute graph, we can represent a document as a keyword vector, and then compute the cosine similarity of two vectors. The resulting value can be treated as the weight of two corresponding document.

Up to this point, we have presented how to calculate the *NodeRank* of a node within one weight graph, and to model a multimedia document collection into different weight graphs corresponding to a number of attributes. Now the remaining problem is how to combine the corresponding *NodeRank*s of a document in different attribute graphs into an overall *NodeRank* as the importance of this document. One straightforward solution is to use a liner combination of these *NodeRank*s:

$$R_n(u) = \frac{\sum_{k=1}^{m}\beta_k \sum_{i=1}^{3}\alpha_i C_i^k(u)}{m} \qquad (2)$$

subject to $\sum_{k=1}^{m}\beta_k = 1$ and $\sum_{i=1}^{3}\alpha_i = 1$

where:

$R_n(u)$: The overall *NodeRank* of document *u*

β_k : The relative importance of the *k* attributes

α_i : The relative importance of the *i* centrality

$C_i^k(u)$: The *i* centrality of node *u* in the *k* attribute graph, calculated by equation (1).

m: the number of attribute graphs, or the number of kind relations between documents.

Different attributes may play different roles in the importance of a document. This can be achieved by assigning varied weights in the above equation.

In the following we formalize our model as Definition 5

Definition 5 *Multi-relation Weight Graph Model (MWGL)*: Given a multimedia collection *D*, where every document $d \in D$ has a set of attributes $\{a_1, a_{2,\ \ldots\ ,} a_m\}$, a series of weight graphs $G_k = (V_k, E_k, \lambda_k)\ (k = 1, \cdots, m)$ is used to represent *m* kind relations based on the different a_k attributes. Within each graph G_k, V_k denotes a set of nodes representing multimedia documents, $E_k \subseteq V_k \times V_k$ is a set of edges indicating the *k* attribute-relation between two documents, and $\lambda_k : E_k \to \Re$, a function assigning weights to the edges. Each edge has a corresponding weight to measure the degree of relation between two documents on a particular attribute.

5. Ranking Multimedia Document Collection in Incorporation with Importance

Based on the previous description, the ranking score of a multimedia document with respect to a given user depends on two factors, namely the importance of a document, and the relevance to the user. It can thereby be calculated in this way:

© The Eurographics Association 2004.

$$R(d) = R_n(d) \times R_s(d) \qquad (3)$$

where $R_n(d)$ is computed by equation (2), and $R_s(d)$ is the relevance value of a multimedia document to a user.

The calculations of $R_s(d)$ are different in current filtering systems. The main techniques include relevance feedback and collaborative filtering, such as, the Rocchio's vector space feedback model [Rob71], and Roberstson's probabilistic networks [Rob77].

We can also write equation (3) in the matrix form as follows:

$$R_{[p\times q]} = R_{s[p\times q]} R_{n[q\times q]} \qquad (4)$$

where

p : The number of users in the system

q: The number of multimedia documents in the collection

$R_{s[p\times q]}$: A user-by-multimedia document matrix, where its entry r_{ij} is the ranking score of the j-th multimedia document for the i-th user.

$R_{n[q\times q]}$: A diagonal matrix consisting of the importance values of multimedia documents in the collection $\begin{bmatrix} R_s^1 & & & \\ & R_s^2 & & \\ & & \cdot & \\ & & & R_s^q \end{bmatrix}$.

The ranking scores of multimedia documents are derived from two parts: the importance value and the relevance value. From the above formulas, it is easy to know the multimedia documents, which play highly important roles in the collection and have high relevance to users, will be presented in high priority orders.

In summary, we have the following theorem.

Theorem 1 Let D be a set of multimedia documents, and a mapping $f : D \to \Re$. Let $\tau \in \Re$ be a positive real threshold value which is between 0 and 1. The following properties of the structure based filtering hold true:

1. $0 \le R(d) \le 1, d \in D$
2. $F_\tau = \{d \mid R(d) \ge \tau \wedge d \in D\}$
3. $F_\tau \subseteq D$
4. $O(d_1) < O(d_2)$, if $R(d_1) \ge R(d_2)$ and $d_1, d_2 \in F_\tau$
5. $O(d) = 0$, if $d \in D \setminus F_\tau$

where

F_τ : A set of remaining multimedia documents corresponding to τ after filtering

$O(d)$: A positive integer indicating the presented order of multimedia document d. If $O(d) = 0$, the multimedia document d will be filtered .

$R(d)$: A ranking score of multimedia document d , which can be calculated by formulae (2) and (3).

6. Algorithm for the Structure-based Filtering

Figure 6 describes an algorithm for computing the *NodeRank*s of documents in a collection using MWGL. This algorithm begins with constructing m kinds of relation graphs from a given document collection D, by expressing each document as a node and each edge as one kind of relation between two documents. Within these m graphs, the *NodeRank*s of each node in each graph is calculated using the equation (1). The importance of each document is then obtained by a liner combination of its *NodeRank* in corresponding graphs with equation (2). Combined with its relevance value with respect to a particular user, it is then ranked in the collection.

Input: *A multimedia document collection* $D = \{d_1, d_2, \ldots, d_n\}$, *a set of document attributes* $A=\{a_1, a_2, \ldots, a_m\}$, *and a set of attribute important weights* $\beta = \{\beta_1, \ldots, \beta_m\}$

Output: *A vector of importance for all documents in D.*

// Construct a series of graph G_k (k=1, …, m)
for k=1 to m
// each document is represented as a node in the k kind relation graph
$V_k = D$
// Build the set of edges E_k
for i =1 *to n*
for j=i-1 *to n*-1
Compute the degree of k kind relation of documents d_i and d_i as the weight of the corresponding edge
end for
end for
end for

for k =1 to *m*
Compute the *NodeRank* vector $\boldsymbol{R}_k$ in graph G_k
end for

$$\boldsymbol{R} = \frac{\sum_{i=1}^{m} \beta_i \boldsymbol{R}_i}{m}$$

Figure 6: The algorithm for calculating importance of documents based on MWGL

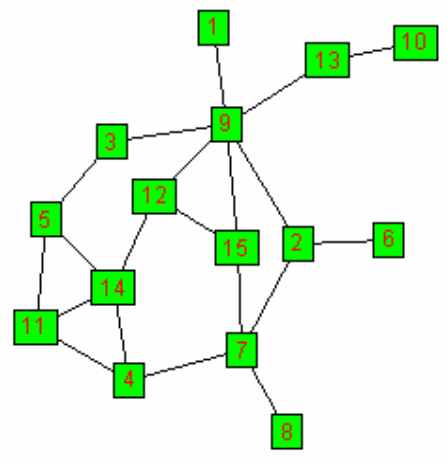
© The Eurographics Association 2004.

7. An Example

In this example, the structure of a small document collection on *receipts* is analysed. For simplicity, we restrict our considerations only to one undirected and connected graph based on the *hyperlink* attribute, and suppose the weight of every edge in the graph is 1. Pages and links of this collection are gathered by using web crawling software named *webCrawler*. The results are shown in Figure 7.

Table 1 illustrates the centrality indices and the ranking scores of the collection. The third column of Table 1 shows the *Betweenness Centrality* indices of the nodes. The *Betweenness Centrality* indices "allow a research to compare different networks with respect to the heterogeneity of the betweenness of the members of the networks" [WSK94]. As shown in Table 1, the *Recipes* node is a prominent node with respect to those measures compared with the other nodes in the collection. From the table, different centrality index may lead to different interpretation. Node 1, for example, has the same *Closeness centrality* value as that of node 11, but with quite different *Betweenness centralities*. As mentioned before, different centrality measures focus on various aspects of structure of a graph. Therefore, we use a combination of them in equation (1) as the importance values, rather than a single centrality measure.

Roughly, there are two kinds of documents in the collection: one is "hub" documents with many links, and another is "sink" documents with incoming links, but without outgoing links [PBM*97] [Kle98]. In Figure 7, for example, *Recipes* (node 9) and *Japanese Fried Rice* (node7) are "Hub" multimedia documents, while *Numerical Recipes* (node1), Oatmeal cookies (Node 6), *Eggs pepper* (node 8) and *Main Dishes* (Node 10) are "sink" multimedia documents. The importance of "Hub" documents' surpasses these of other documents so that they have relatively high importance values.



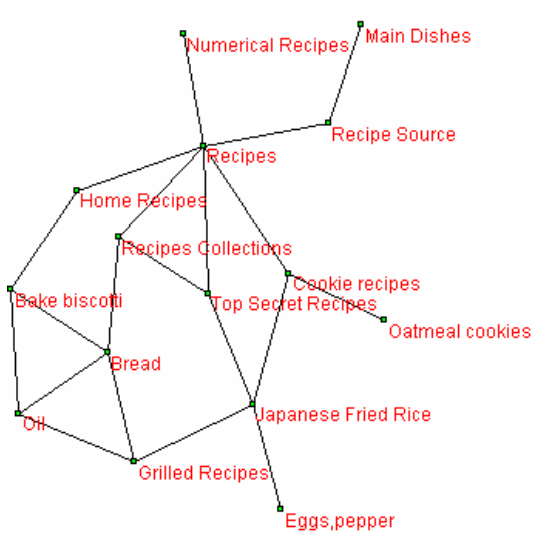


Figure 7: The structure of a small document collection on "recipes"

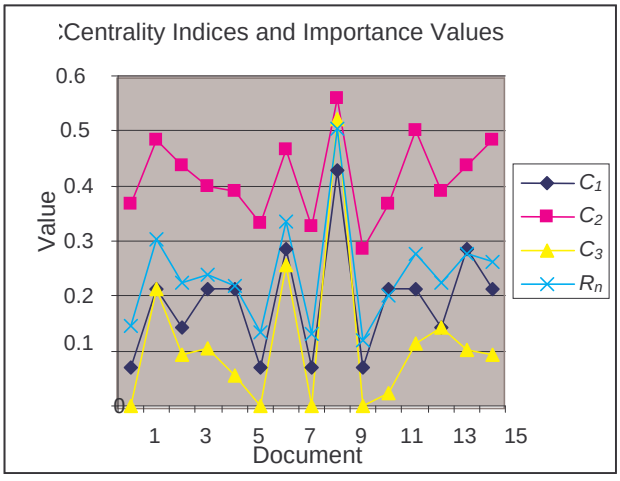


Figure 8: Comparisons of Centrality indices and importance values.

"Hub" is a transitional document through which users move to certain destinations, while "sink" tends to be a final destination.

Figure 8 illustrates the centrality indices and importance values of documents in the collection.

Suppose there are two users, and their relevance vectors, i.e. the user relevant values of the documents, are R_{s1} and R_{s2} as shown in Table 1. We can then construct the user-by-document matrix $\boldsymbol{R_s}$ and the importance value of a diagonal matrix $\boldsymbol{R_n}$. Therefore, the ranking scores of documents are

© The Eurographics Association 2004.

Node	C_1	C_2	C_3	R_n	R_{s1}	R_{s2}	R_1	R_2	O_1	O_2
1	0.071	0.368	0.000	0.146	0.7271	0.9797	0.1062	0.1430	9	4
2	0.214	0.483	0.212	0.303	0.3093	0.2714	0.0937	0.0822	11	9
3	0.143	0.438	0.093	0.225	0.8385	0.2523	0.1887	0.0568	4	11
4	0.214	0.400	0.104	0.239	0.5681	0.8757	0.1358	0.2093	7	2
5	0.214	0.389	0.055	0.219	0.3704	0.7373	0.0811	0.1615	12	3
6	0.071	0.333	0.000	0.135	0.7027	0.1365	0.0949	0.0184	10	13
7	0.286	0.467	0.255	0.336	0.5466	0.0118	0.1837	0.0040	5	15
8	0.071	0.326	0.000	0.132	0.4449	0.8939	0.0587	0.1180	14	6
9	0.429	0.560	0.522	0.504	0.6946	0.1991	0.3501	0.1003	1	8
10	0.071	0.286	0.000	0.119	0.6213	0.2987	0.0739	0.0355	13	12
11	0.214	0.368	0.022	0.201	0.7948	0.6614	0.1598	0.1329	6	5
12	0.214	0.500	0.114	0.276	0.9568	0.2844	0.2641	0.0785	2	10
13	0.143	0.389	0.143	0.225	0.5226	0.4692	0.1176	0.1056	8	7
14	0.286	0.438	0.103	0.276	0.8801	0.0648	0.2429	0.0179	3	14
15	0.214	0.483	0.092	0.263	0.1730	0.9883	0.0455	0.2599	15	1

Table 1: Parameters in the Collection

calculated according to formula (3):

$$R = R_s R_n$$

$$= \begin{bmatrix} 0.7271 & 0.3093 & \cdot & \cdot & 0.1730 \\ 0.9797 & 0.2714 & \cdot & \cdot & 0.9883 \end{bmatrix} \begin{bmatrix} 0.1460 & & & & \\ & 0.3030 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & 0.2630 \end{bmatrix}$$

$$= \begin{bmatrix} 0.1062 & 0.0937 & \cdot & \cdot & 0.0455 \\ 0.1430 & 0.0822 & \cdot & \cdot & 0.2599 \end{bmatrix}$$

The results are also shown in Table 1. Note that the order number for document 9 to be presented to the user 2 is only 8, although it has the highest importance value in the collection.

For the user 1, if the threshold $\tau = 0.06$ is chosen, then documents 15 and 8 will not be presented. O_1 and O_2 in Table 1 also give the ranking order of the presentation of documents to users 1 and 2, respectively.

8. Related Work

There exist many information filtering systems. The main mechanisms of these systems involve three problems: how to represent a user's information (query or profile) and the multimedia document set for an effective comparison; how to compare the above representations? How to use the feedback mechanism to improve the performance of systems? Our approach focuses on linking the multimedia document collection to the users not for a comparison, but for more accurately representing every user's need. Equally, it is to access the important and relevant information. Actually our approach models common features among diverse user profiles.

The Information Lens system [MGR97] generates rules based on the structure of a mail message to filter mails. However, the extracted structure is within a document. Other link analysis of the structure of information includes HITS and PageRank algorithms [Kle98][PBM*98], but they use the link structure to improve web search engines.

Our approach differs from other approaches in that it combines the importance of a multimedia document into their relevance as part of user profiles. The proposed approach explores the roles of multimedia documents, regardless of the content of multimedia documents. Our approach can efficiently filter the multimedia documents.

© The Eurographics Association 2004.

9. Conclusion

The presentation of user profiles is an important issue in information filtering systems. Although different users have diversified profiles, this paper has presented a new framework for information filtering. With this framework, we have described a new approach to determining the importance values of multimedia documents in the collection to form part of user profiles, on the assumption that every user needs both important and relevant multimedia documents. This approach extends the concept of centrality used in social network analysis to explore different roles of multimedia documents, and then gives overall ranking scores of multimedia documents along with relevant values. Our approach explicitly takes advantage of the link structure of multimedia documents. It thus does not depend on the contents of multimedia documents. The future work will explore more applications.

Reference

[BC92] BELKIN N. J. AND CROFT W.B.: Information filtering and information retrieval two sides of the same coin? *Communications of the ACM 35*, 12 (1992), 29-38.

[Fre78] FREEMAN L.C.: Centrality in social networks: I. Conceptual clarification. *Social Networks*, 1(1978), 215-39, Cambridge University Press.

[Kle98] KLEINBERG J.: Authoritative Sources in a Hyperlinked Environment, *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[HSS01] HANANI U., SHAPIRA B. AND SHOVAL P.: Information Filtering: Overview of issues, research and systems, *User Modelling and User-Adapted Interaction* 11 (2001), 203-259.

[Lin94] LINDA S.: Relevance and Information Behaviour. *Annual Review of Information Science and Technology* (ARIST) 29(1994), 3-48.

[Mal87] MALONE T.W..: Intelligent Information-Sharing Systems, *Communications of the ACM* 30, 5(1987), 390-402.

[MGR97] MALONE T.W., GRANT K. R., RAO R.: Semi structured messages are surprisingly useful for computer-supported coordination. *ACM Trans. Off. Inf. Syst.* 5, 2 (1987), 115-131.

[MMP*97] MOSTAFA J., MUKHOPADHYAY S., PALAKAL M., LAMW. : A Multilevel Approach to Intelligent Information Filtering: Model, System, and Evaluation, *ACM Transactions on Information Systems* 15, 4 (October, 1997), 368–399.

[Oar97] OARD W.D.: The state of the art in text filtering. *User Modeling and User Adapted Interaction* (UMUAI) 7, 3(1997), 141-178.

[PBM*98] PAGE L., BRIN S., MOTWANI R., AND WINOGRAD T.: The PageRank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project, 1998.

[Roc71] ROCCHIO J.: Relevance feedback in information retrieval. (1971), 313–323.

[Rob77] ROBERTSON S. F.: The probability ranking principle. *Journal Of Multimedia documentation* (1977), 294–304.

[SJ00] SCOTT, JOHN: *Social Network Analysis*: *A Handbook, London*: (2000), Sage.

[WSK94] WASSERMAN, STANLEY and KATHERINE F.: *Social Network Analysis: Methods and applications* (1994)

© The Eurographics Association 2004.