APPLYING CROSS-TOPIC RELATIONSHIPS TO INCREMENTAL RELEVANCE FEEDBACK

Terry C H Lai, Stephen C F Chan, Korris F L Chung
Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China

Keywords: Internet Information Retrieval, Search Topics, Bayesian Networks

Abstract: General purpose search engines like Google and Yahoo define search topics for the purpose of document organization, yet their hierarchical structures cover only a portion of topic relationships. Search effectiveness can be improved by using search topic networks, in which topics are linked through semantic relations. In our search model, is-child and is-neighbor relations are defined as relations among search topics, which in turn can serve as search techniques; the *is-child* relation allows searching from general concepts, while the *is-neighbor* relation provides fresh information that can help users to identify search areas. This search model uses the Bayesian Networks and the incremental relevance feedback. Our experiments show that search models using the Bayesian Networks and the incremental relevance feedback improve search effectiveness.

1 INTRODUCTION

Web-based personalization, for example, commercial sites advertising products based on users' buying patterns, has greatly complicated the work of search engines and other web applications. The Outride Personalized Search System (Pitkow, 2002), for example, uses information such as content interests, click streams, and search histories to provide a personalized query result. Client-based architectures such as the CI Spider and the Meta Spider (Chau, 2001) provide further personalized analyses. It is a complicated matter to personalize searches and one of the most difficult steps is the prediction of users' search purposes. Even though search engines may maintain a user's history, it remains uncertain which part of the history is applicable to a particular search task. In any case, it is almost impossible to keep all personal information because of heavy demands on processing time (Jeh, 2003). These problems can be effectively addressed using incremental relevance feedback. This approach searches without depending on a user's history heavily, using users' feedback to make the result list of the next search round more relevant. Ingwersen (Ingwersen, 1992) has similarly noted the importance of taking into account situational factors like users' feedback in which, after reading documents returned by a search engine, readers will choose certain documents that are used to modify an original query for the next round of search.

Previous works in the area of incremental relevance feedback (Salton, 1990) has focused on search effectiveness reflected by precision and recall in experiments, always assuming that users would identify the relevance of documents. However, it is unlikely that users are in fact willing to do so much extra work when searching. Therefore, White (White, 2002) developed an implicit relevance feedback system without users' explicit involvement. Experiments showed that the result of that implicit approach is quite positive. Allan (Allan, 1996) found that the number of judgments has a proportional effect on precision. But it is difficult for Internet searchers to make large numbers of judgments. Iwayama (Iwayama, 2000), on the other hand, found that a decrease in the number of judgments produced a drop in search effectiveness of only 10%. Clearly then, acceptable results can still be kept with a small number of relevance judgments. Thus, the incremental relevance feedback approach is found to be feasible and useful.

Jansen (Jansen, 2000) found that 62% of queries submitted to the Excite search engine contained only one or two terms. Users combining this technique with ambiguous keywords are unlikely to find the documents they want quickly. This situation improves, however, if they provide a search area as well. Well-defined search areas filter out confusing

C H Lai T., C F Chan S. and F L Chung K. (2004).
APPLYING CROSS-TOPIC RELATIONSHIPS TO INCREMENTAL RELEVANCE FEEDBACK.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 356-363
DOI: 10.5220/0002620203560363
Copyright © SciTePress

and unwanted materials so users can find their targets more quickly. Many general purpose search engines have classified their documents into different search topics that can be defined either by human experts or algorithms (Liu, 2003). In fact, much work has focused on the search topic. A personalized search system (Liu, 2002) used search topics to clarify ambiguous queries. Experiments (Kelly, 2002) show that strong familiarity with a search topic increases search effectiveness.

The hierarchical structures of search engines contain over a thousand search topics. As a result, users' choices of topics are not usually the most appropriate for a particular query, but are merely related topics. Relationships among topics should be useful for seeking information in conditions of uncertainty. Our study of the relationships among topics, which surveyed the patterns of cross-topic hyperlinks in the Yahoo search engine, has revealed that Yahoo's hierarchical structure covers only some topic relationships. Our search model defines *is-child* and *is-neighbor* relations in a way that it studies relationships among topics. An *is-child* relation is a kind of relation between child and parent topics, whereas an *is-neighbor* relation represents all other topic relations. These two relationships can be useful in different search strategies. *Is-child* relations connecting child and parent topics help users who start searching with only a general idea. On the other hand, *is-neighbor* relations providing connections to some related topics act as an innovative pathway for brain storming and browsing. In our search model, the Bayesian Networks (Jensen, 1999) that encompass both *is-child* and *is-neighbor* relations are used to determine the search scope and the document ranking. Moreover, we use the incremental relevance feedback technique to modify user queries. Experiments show this model requires fewer result rounds to reach the most relevant information.

In Section 2, we discuss the cross-referencing of topics in Yahoo. Section 3 describes our search model using the Bayesian Networks and the incremental relevance feedback. Section 4 presents experiments, and Section 5, discussion. Section 6 summarizes our work and outlines our future work.

2 PATTERN OF CROSS REFERENCES

To understand the cross referencing of general purpose search engines, we study a subset of Yahoo. Documents inside Yahoo are organized according to a hierarchical topic structure with each document assigned to a topic on the basis of its contents. For this study, we chose the branch "Computer & Internet" (1572 topics and 970083 web pages). Relationships between the topics are studied by counting the number of hyperlinks of web pages across the topics. The terms "Link" and "Reference", used interchangeably in the following discussion, refer to the connection between web pages.

2.1 Interrelated topics

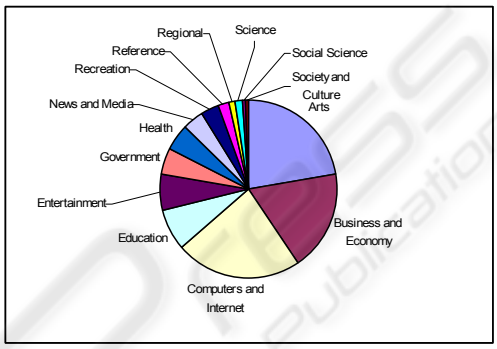

Figure 1: Reference distribution of the "Computers & Internet" Branch in Yahoo

Topics are interrelated. Figure 1 illustrates the link distribution from the "Computers & Internet" branch to all the other branches. We found that web pages in the "Computer & Internet" branch have references pointing to all other branches. From the figure, it is clear that some relationships between branches are much stronger than others. For example, the relationship between "Computers & Internet" and "Arts" and "Business & Economy" are the strongest. The differences in the strength of the relationships among branches motivate our further studies of relationships of branches' topics.

2.2 Overall distribution of references

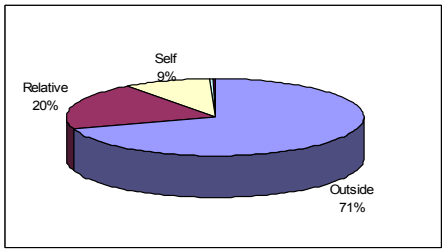

Figure 2: Type of Reference in the "Computer & Internet" branch

We classified references into different types; "Ancestor" represents the links from a document in a descendant topic to a document in an ancestor topic under their hierarchical structures.

"Descendant" is the opposite of "Ancestor". "Self" means that the links point to the same topic. "Relative" is a link pointing to a relative topic. "Outside" means the link connects two topics that are not related in the above manners. The result of our classification is shown in figure 2. We can see that "Relative" and "Outside", with high proportions, are quite significant in the figure. More importantly, these types of relationship haven't been considered a little by the hierarchical structures such as Yahoo's. Furthermore, it is surprising that "Ancestor" and "Descendant" are not so common. That means web pages are less likely to contain hyperlinks pointing to those web pages in their ancestor and descendant topics. This may be a consequence of the fact that hyperlinks represent such a wide range of semantics. This finding arouses the study of relations not in the hierarchy.

2.3 Reference distribution by topic

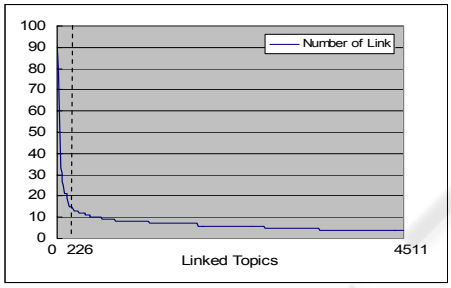

Figure 3: Reference distribution of "News and Media" from "Computers and Internet" in Yahoo

Figure 3 illustrates the reference distribution of the subject topic "News and Media" in the "Computer & Internet" branch. The x axis of this figure represents all search topics connected to "News and Media" through cross references of hyperlinks, whereas the y axis is the number of references between the subject topic ("News and Media") and its connected topics. The total number of connected topics is 4511. The figure indicates that only a few topics have a high number of references linking to the subject topic. The dotted line in the figure indicates 5% of all connected topics. Similar result is obtained from most topics. Probability based model is used to represent the uneven distributions mirrored in these findings.

3 SEARCH MODEL USING SEARCH TOPIC NETWORKS

Unlike traditional relevance feedback techniques that use only document contents to improve search effectiveness. Our search model also takes documents' topics into account. This extra information is useful in defining a search scope and in ranking retrieved documents.

3.1 Search topic networks

In our proposed system, relations among these search topics used to build search topic networks, *ST Nets*, are in the form of the Bayesian Networks, a powerful tool for knowledge representation and reasoning in conditions of uncertainty (Pearl, 1998). Since our pervious observations show that some topic-topic relations are much stronger than others, probability-based model seems to be a good way to represent such uneven distributions. *ST Nets* are composed of search topics connected by links, where a link between topics represents either an *is-child* relation or an *is-neighbor* relation. An *Is-child* relation links a parent topic to a child topic, whereas an *is-neighbor* relation links relative topics or topics in different branches. Some relationships like "Relative" links or "Outside" links are not fully included in hierarchical structures such as Yahoo's. To cover these hidden relationships, our search model has an *is-neighbor* relation that links relative topics or topics in different branches. A link among relative topics refers to a connection of any topics within a same branch, except the connection between parent and child topics.

The strength of an *Is-child* relation is estimated based on the number of web pages within a topic. Suppose topic A.1 has A.1.1 and A.1.2 as its child topics, A.1.1 contains five web pages, whereas A.1.2 contains two web pages. Assuming that web pages in child topics can also be grouped in their parent topics, A.1 contains seven web pages by summing up the number of web pages in its child topics. The weight $P(A.1.1 \mid A.1)$ of link from A.1 to A.1.1 is 2/7, representing the probability of getting relevant result in a more specific topic A.1.1 with a more general topic A.1 as a starting point. In reality, what the users are looking for is in a specific topic but they start searching from a more general topic.

The strength of an *is-neighbor* relation between topics, on the other hand, is calculated by counting the number of hyperlink connecting web pages within these topics. And each link between topics is weighted in the form of conditional probability. The

higher is the weight of the link, the stronger are the relationship between the topics. Figure 4a & b illustrate the way to calculate the weight of the *is-neighbor* relation between topic P and Q, provided that topic P and Q are either relative topics or topics in different branches. There are five links in P and one link between P and Q. Thus, the weight of the *is-neighbor* relation between P and Q $P(Q \mid P)$ is 1/5 (equation 1).This value can be considered the probability that a related topic (e.g., Topic Q) contains relevant information given a user's selected topic (e.g., Topic P) in our search model. Very often users learn from many channels, and may not always start searching at the most appropriate topics. For example, papers on data mining may not be found exclusively in data mining journals. Instead, they may also be found in information retrieval journals, or database journals, or even mathematics journals. It is not surprising that non-expert users may choose related but not the most relevant topics at the beginning of search. In this case an *is-neighbor* relation that provides alterative suggestions is useful. Finally, Figure 5 shows a sample *ST Net* modeled in form of a Bayesian Network, where all topics are connected by weighted links. In fact, some information retrieval systems (Popescul, 2001) also use approaches of Bayesian networks to select documents.

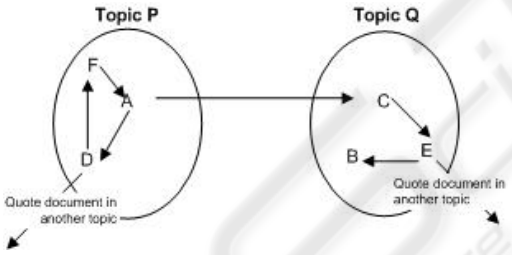

Figure 4a: Link Structure of P and Q

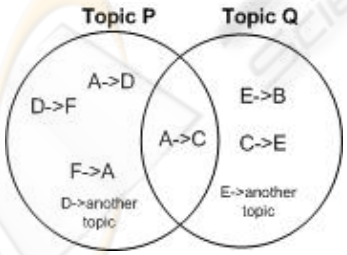

Figure 4b: Conditional Probability

$$P(Q \mid P) = \frac{\text{Number of link between P and Q}}{\text{Number of link in P}} \text{ (Equation 1)}$$

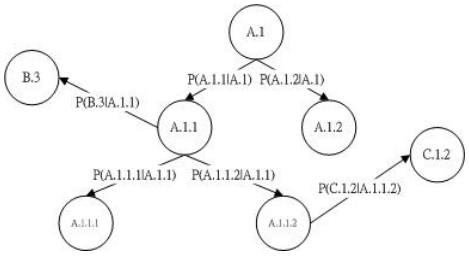

Figure 5: Search Topic Networks

3.2 Search model using incremental relevantal feedback and search topic networks

Our search model makes use of incremental relevance feedback and *ST Net* to improve the accuracy of search results. In each round of result list, users have to judge those retrieved documents. Suppose topics of these documents are well defined. (Many search engines have classified their documents into different topics.) The topics of those relevant documents are used act as initial search topics and are used to find related topics by using *ST Nets* described before. Referring to Figure 5, we suppose an initial topic of a relevant document is A.1.1. Potentially related topics are those topics that are linked from the given topic directly or indirectly by an *is-child* relation, or connected by an *is-neighbor* relation directly. In this case, we have A.1.1.1, A.1.1.2, and B.3 as potentially related topics of A.1.1.

Unlike an *is*-child relation, an *Is-neighbor* relation isn't generally a transitive relation. In a case study (figure 6), we have studied 22686 relations among topics. The actual relation strength refers to the probability of one topic given another topic, that is, the way we find the weight of *is-neighbor* relation. A projected relation strength is calculated by multiplying the weights of topic relations where the object topic of one relation is the subject topic of another one, e.g., R_1 (A→B) and R_2 (B→C), intuitively what the strength of A→C might be if *is-neighbor* relations are transitive. The figure shows that the actual relation strength doesn't change significantly no matter how large the projected relation strength is. Thus, an *is-neighbor* relation is less unlikely to transitive according to the behavior of hyperlink patterns. Therefore, if B.3 further links another topic (e.g., topic D.2), D.2 is not supposed to be a related topic of A.1.1. However, if A.1.1.1 has another descendant (e.g., A.1.1.1.1), this descendant topic is also considered as a candidate of related topic of A.1.1.

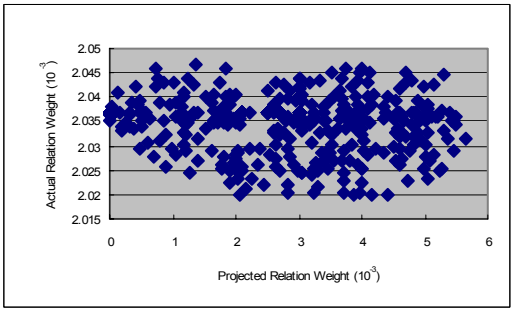

Figure 6: Transitive property of topic relation

We start calculating how related a topic is to a given topic with the weight of the topic relationship. Equation 2 calculates the strength of topic Y with respect to selected topic X. Suppose there are m topics (N_1, N_2… N_m) between X and Y. In current search model, we only consider two kinds of connection: topics can either connect together with an *is-child* relation, or topics are linked by an initial topic with an *is-neighbor* relation directly. Whichever type of connection is calculated, the strength of Y w.r.t. X ($S_{x\to y}$) is calculated by using the joint probability of all topics involving the chain from X to Y. By using the chain rule of the Bayesian Networks, the strength of Y w.r.t. X is calculated by equation 2. Obviously, *is-child* and *is-neighbor* relations have different definitions. To compare the closeness of two topics that are connected to one topic with an *is-child* and an *is-neighbor* relation, a normalization factor α, β are used to normalize the weight of *is-child* and *is-neighbor* relation respectively. Therefore, the new weight of relation should be calculated by multiplying its original weight (the weight of an *is-child* relation or an *is-neighbor* relation) by corresponding normalization factor (α or β). In our current search model, we ignore this consideration by choosing both normalization factor α and β = 1. However, these factors reflect search orientations of users. For those users who have already had some general ideas about their search targets, the normalization factor of an *is-child* relation α should be higher than β, because they want to get more information that may exists in more specific topics. On the other hand, for those users who do not have a prior knowledge about their search targets, the normalization factor of an *is-neighbor* relation β should be higher so as to return some related and innovative topics for browsing.

As mentioned, suppose a user has chosen a document in topic A.1 as a relevant document, and three related topics (A.1.1.1, A.1.1.2, and B.3) are found with their strengths with respect to topic A.1.1. For the next iteration of the search, only those documents in these related topics are considered. Moreover, our search model has employed the incremental relevant feedback technique to modify an original query. Relevance feedback algorithms, such as, F4, Rocchio (Rocchio, 1971), Taylor, make use of users' feedback to modify an original query. The modified query is then used for the next round of search within a scope of the related topics. These steps are done continuously until the user is satisfied with a search result. Retrieved documents are then ranked by using equation 3, which takes the document's contents and the search topics into account. *RD* is a document ranking function. It chooses the cosine similarity function, HITS, PageRank (Page, 1998) or some other method, depending on properties of the collection. *RC*, is a ranking function for topics, ranks the topics based on how related they are to selected topics. The way to determine the strength of relationship between topics has been mentioned in previous sections. RC just ranks topics whose have larger strength in a higher position. Finally, documents are ranked by using equation 3 in an ascending order of this *Rank* function. In our experiments, we have chosen γ=2 and λ = 1 through some case studies. Parameter γ considers the importance of document content when ranking, whereas parameter λ considers the importance of the search topic of corresponding document.

$$S_{X\to Y}=\begin{cases}P(X)\alpha P(N_1\mid X)\alpha P(Y\mid N_m)\prod_{i=2}^{m}\alpha P(N_i\mid N_{i-1}) & \text{in the case of } \textit{is-child}\\ P(X)\beta P(Y\mid X) & \text{in the case of } \textit{is-neighbor}\end{cases} \quad \text{(Equation 2)}$$

where $P(X)=\frac{\text{Number of document in X}}{\text{Number of document in database}}$, N_i is the topic in the chain of topic between the topic *X* and *Y*, *m* is the total number of topic between *X* and *Y*, $P(N_i|N_{i-1})$ is either the weight of an is-child relation, or an *is-neighbor* relation.

$$Rank_{d1,c1}=\frac{1}{\gamma}RD(d1)+\frac{1}{\lambda}RC(c1) \qquad \text{(Equation 3)}$$

Where d1 and c1 refer to a document and a corresponding topic

4 APPLICATIONS TO SEARCH TOPIC NETWORKS

To validate the improvements to the search model due to the use of *ST Nets*, two search systems have been set up. One uses incremental relevance feedback alone, whereas the other uses both the incremental relevance feedback supplemented by *ST Nets*. Yahoo acts as an underlying search engine in both cases. For the incremental relevance feedback, a term frequency (Equation 4) is used to measure the importance of a term in documents. The Rocchio method (Rocchio, 1971) is chosen for query

modification. This method (Equation 5) takes those important terms from an original query and a relevant document set for query expansion. It also gets rid of those terms from an irrelevant document set. In our experiment, we choose 1, 3, and 0 as the values of parameter θ, σ, and ϕ respectively through case studies. Some previous experiments (Aalbersberg, 1992) show that considering those terms in an irrelevant document set doesn't greatly affect accuracy. Moreover, these extra processes may increase the waiting time. Therefore, the factor ϕ is set to 0.

$$W_{i,d} = \frac{\text{freq}_{i,d}}{\max \text{freq}_{i,D}} \qquad \text{(Equation 4)}$$

where $W_{i,d}$ refers to the weight of the term i in a document d, $freq_{w,i}$ is the frequency of the term i in a document d, a max $freq_{i,D}$ is the maximum frequency of the term i in a document set D

$$Q^{new} = \theta Q^{old} + \sigma \frac{1}{|D|} \sum_{D} d_i - \phi \frac{1}{|D'|} \sum_{D'} d_j \qquad \text{(Equation 5)}$$

where Q^{old} is a keyword set of initial query, D is a relevant document set, D' is an irrelevant document set, a document d_i ,d_j is a vector of term, Q^{new} is a new keyword set calculated by this equation.

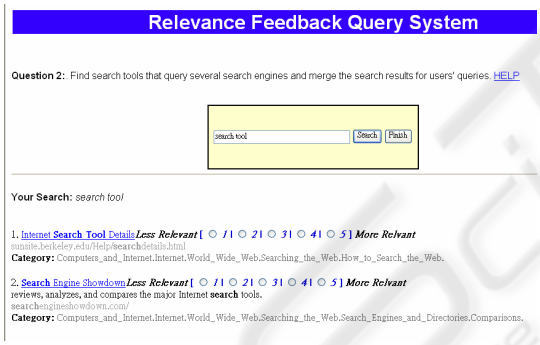

Figure 7: User Interface of Relevance Feed Query Svstem

Figure 7 shows the interface of our experimental system. Two search systems share the same interface except the search algorithms. To determine the relevance of a document, we have invited 5 subjects to take the experiments. All of them are familiar with using search engines. Subjects have to finish all tasks listed in Table 1 with our search systems, and each question has been provided some keywords as an initial query. All these keywords are too general for users to reach their search targets directly, so users have to select those relevant documents from the result list in each search round, progressively modifying an original query to get a better result. Users are required to judge each retrieved document on a scale of 1 to 5, where 1 is the least relevant and 5 is the most relevant. The top 5 most relevant documents are chosen randomly from the result list. These five documents are used to modify the original query. In each round of the incremental search, this search system takes a maximum of the first 25 documents from the Yahoo search engine because users are more likely to pay more attention to those documents on the first result page that usually doesn't contain more than 25 documents.

We have another search model, on the other hand, uses *ST Nets* to supplement the incremental relevance feedback. This approach takes both contents and topics of relevant documents as users' feedback. The contents of relevant documents used to modify a users' queries, whereas the topics of relevant documents are used to find their most related topics through *ST Nets* described in Section 3. As mentioned, related topics chosen from *ST Nets* are used as a search scope for further processing. Thus, the system submits those modified queries and takes a maximum of fifty documents that belong to the related topics from *ST Nets* through the Yahoo search engine. This process filters out those documents from irrelevant topics. By using equation 3, these retrieved documents are re-ranked. RD in equation 3 refers to the yahoo's original document ranking algorithm in our experiments. If the number of retrieved document exceeds 25, only the first 25 documents will be sent to users for judging. Again, users have to accomplish those tasks listed in Table 1 by using this system, and award a mark for each retrieved document.

Table 1: Question Table with initial query

	Task	Initial Query
T1	Find some viruses that attack web servers. Web pages should include descriptions of and solution for the viruses.	Virus
T2	Find search tools that query several search engines and merge the search results for users' queries.	Search Tool
T3	Find some information about internet security, such as, service providers, techniques, and internet security software.	Internet Security
T4	Find programming tools for palm software development.	PDA Program
T5	Find some information of a laser color printer for printing a digital photo, such as cost or functions.	Printer
T6	Find desktop customization software that personalizes your wallpapers, icons, mouse cursors, etc.	Desktop Theme
T7	Find information about home network for file and printer sharing, interest access, etc.	Home Network
T8	Find magazines that introduce the latest web programming information	Program Magazine

5 DISCUSSION

The result obtained from the search models with and without *ST Nets*.

Figure 8 shows the similarity coefficient of result lists returned by our two search models, as calculated by equation 6. The Higher similarity is the coefficient, the closer are the two result lists. The first round of result is the same for both models, as these two search models take the same pre-defined initial query for each task. Result lists in round two shows the greatest difference. However, they become more similar again in the last round where the most relevant result is obtained.

$$Similarity(r_1, r_2) = \frac{\text{number of document existing in both lists}}{\text{Total number of documents}} \quad \text{(Equation 6)}$$

where r_1 and r_2 represents two different result lists

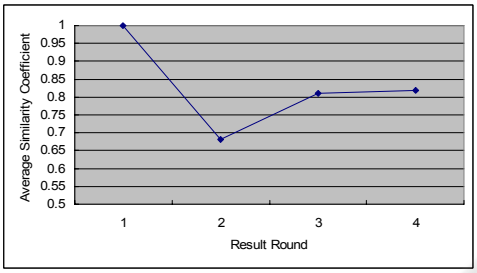

Figure 8: Average Similarity Coefficient of result lists between two search models

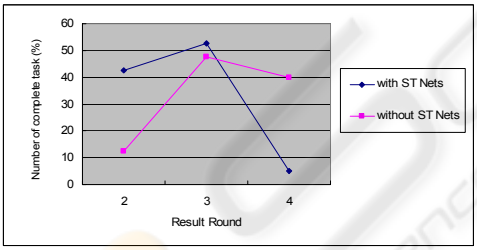

Figure 9: Number of result rounds to use to reach a precision peak

Figure 9 shows how many result round of search is used to obtain the best result list, in which most relevant documents are retrieved. The line in figure 9 illustrates the percentage of search task completed within any particular result round by the approach it represents. The search model with *ST Nets* usually takes fewer result rounds to complete the tasks. Users using *ST Nets* usually take 2-3 result rounds to reach the result lists with the highest precision. On the other hand, system without *ST Nets* requires 3-4 result rounds to get the best results. This shows that considering topics of documents can shorten the number of result rounds for compose the best result list.

Figure 10 shows the quality of the result lists returned by the search systems. This figure illustrates the average precision of result lists in each round of search. The precision of result list is calculated by taking the number of relevant documents over the total number of retrieved documents. A document is said to be relevant if they are marked as 4 or 5. In fact, the precision peak of both systems is more or less the same. However, the *ST Net* approach tends to give better result lists in earlier rounds.

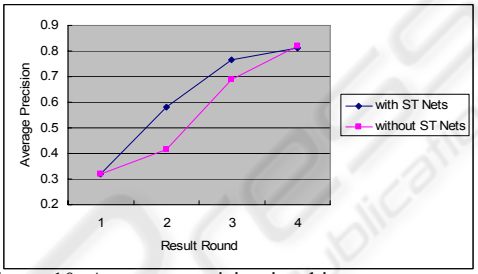

Figure 10: Average precision is taking an average value of all precision values of result lists in each round of

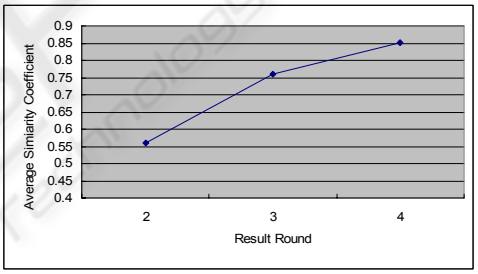

Figure 11: Average similarity of search topics visited by two search models search.

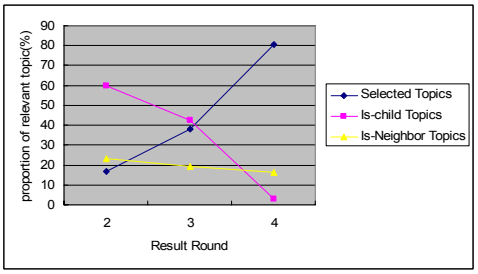

Figure 12: Proportion of topics providing relevant documents.

Is-child and *is-neighbor* relations are useful for information seeking in the *ST Net* approach. Figure 11 illustrates the similarity of topics in the result lists generated by both search systems. The search topics involved in result lists of both models become more similar as the last rounds are approached. Figure 12 shows the proportion of relevant documents within following categories: a selected topic, an *is-child*

topic, and an *is-neighbor* topic. Selected topics are those topics whose documents are judged as relevant information by users in the previous search round. *Is-child* and *is-neighbor* topics are those topics that are connected by an *is-child* relation and an *is-neighbor* relation respectively with selected topics. Again, a document is said to be a relevant document if users award it a 4 or a 5. Figure 12 shows that the proportion of *is-child* topic drops gradually. This probably means that once users have reached suitable topics through a few rounds, any further specific child topics are less important to users. A continuous increase in the proportion of target topics reflects the fact that relevant feedback improves search quality. Moreover, there are some relevant documents belong to *is-neighbor* topics in each result round. That means *is-neighbor* topics have some contributions for finding relevant information.

6 CONCLUSIONS

In this paper, we discuss the pattern of cross-topic reference in Yahoo. The findings show that topic relationships are not limited to a parent-child relation. Instead, there are some significant and useful relationships among relative topics or topics from different branches. *Is-child* and *is-neighbor* relations have been defined to represent those relationships among the topics, and *ST Nets* are formed based on such relations. Experiments show that users using the incremental relevant feedback and *ST Nets* are able to get the most relevant result with fewer result rounds. In future, *ST Nets* will be further investigated. Now we only consider those topics that are all connected by *is-child* relations, or are linked by an *is-neighbor* relation directly. However, there are more types of combinations of *is-child* and *is-neighbor* relations that should bring some meaningful information for different situations. In addition, we are trying to apply *ST Nets* to different information retrieval systems in the fact that topic relationships should have a wide range of usage for information retrieval.

ACKNOWLEDGEMENTS

The research reported in this paper was partially supported by the Hong Kong Polytechnic University Research Grant A-PE35.

REFERENCES

Aalbersberg, I. J., 1992. Incremental Relevance Feedback,15th Ann Int'I SIGIR '92, Denmark, pp. 11-22.

Allan, J., 1996, Incremental Relevance Feedback for Information Filtering, SIGIR'96, Zurich, pp. 270-278.

Chau, M., Zeng, D., Chen, H., 2001. Personalized Spiders for Web Search and Analysis, JCDL '01, Virginia, USA, 24-28 Jun, pp. 79-87

Ingwersen, P., 1992. Information Retrieval Interaction, Taylor-Graham.

Iwayama, M., 2000. Relevance Feedback with a Small Number of Relevance Judgements: Incremental Feedback vs. Document Clustering, SIGIR'00, Athens, Greece, July, pp. 10-16.

Jansen, B. J., Spink, A., Saracevic, T., 2000. A study and analysis of users queries on the Web, Information Processing & Management 36. 2., pp. 207-227.

Jeh, G., Widom, J., 2003. Scaling Personalization Web Search, WWW2003, Hungary, 20-24 May, pp. 271-279.

Jensen F. V., 1999. The book, Bayesian Networks and Decision Graph, Springer.

Kelly, D., Cool, C., 2002. The Effects of Topic Familiarity on Information Search Behavior, JCDL '02, Portland, Oregon, USA, 13-17 Jul, pp. 74-75.

Liu, B., Chin, C. W., Ng, H. T., 2003. Mining Topic-Specific Concepts and Definitions on the Web, WWW 2003, Budapest, Hungary, 20-24 May, pp. 251-260.

Liu, F., Yu, C., Meng, W., 2002. Personalized Web Search by Mapping User Queries to Categories, CIKM '02, Virginia, USA, 4-9 Nov, pp. 558-565.

Page, L., Brin, S., Motwani, R., Winograd, T., 1998. The PageRank citation ranking: Bringing order to the Web. Technical report, Standford University Database Group, http://citeseer.nj.nec.com/368196.html.

Pearl J., 1998. Probabilistic Reasoning in Intelligent Systems: networks of plausible inference, Morgan Kaufmann Publishers, San Mateo, CA.

Pitkow, J., Shcutze, H., Cass, T., Cooley, R., 2002. Personalized Search, Communications of the ACM, Vol. 45, No. 9, pp.50-55.

Popescul, A., Ungar, L. H, Pennock, D. M., Lawrence, S., 2001. Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Date Environment, UAI, Seattle, USA, 2-5 Aug 2-5, pp.437-444.

Rocchio J. J., 1971. Relevance Feedback in Information Retrieval, Prentice-Hall Incorporation.

Salton, G., and Buckley, C., 1990. Improving Retrieval Performance by Relevance Feedback, Journal of the America Society for Information Science, 41(4), pp. 288-297.

White, R. W., Ruthven, I., Jose, J. M., 2002. The use of implicit evidence for relevance feedback in web retrieval, European Conference on Information Retrieval Research, UK, March, pp. 93-109.