User Search Goal as Document with Feedback Session Using Clustering

Nithya BP [#1], Sreerekha B[#2]
[#]*CSE Department, Kannur university*

Abstract— In web search applications, queries are submitted to search engines to represent the information needs of users. Sometimes queries may not exactly represent users' specific information needs since many ambiguous queries may cover a broad topic and different users may want to get information on different aspects when they submit the same query. User search goals can be considered as the clusters of information needs for a query. The feedback session is defined as the series of both clicked and unclicked URLs and ends with the last URL that was clicked in a session from user click-through logs. Then, propose a novel optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs. Then cluster these pseudo-documents to infer user search goals and depict them with some keywords. This paper proposed bisecting k mean clustering to cluster the pseudo-documents and a summarization technique to highlight the core information for the user to understand the information easily.

***Keywords*— User search goal, feedback session, Pseudo-Documents, Summarization, Evaluation Criterion**

I. INTRODUCTION

Data mining is a process of extraction of hidden predictive information from large databases. It help the companies to focus on the most important information in their data warehouses. Data mining tool help the businesses to make proactive.

In web search applications, queries are submitted to search engines to represent the information needs of users. Sometimes queries may not represent users' specific information needs since many ambiguous queries may cover a broad topic and different users may want to get information on different aspects when they submit the same query. For example,the query "the sun" is submitted to a search engine. Some user want to learn the natural knowledge of solar system and some other want to locate the homepage of a United Kingdom newspaper. So it is necessary to capture different user search goals in information retrieval.

User search goals are defined as the information on different aspects of a query that user groups want to obtain. So user search goals are clusters of information needs for a query. This paper proposed a novel approach to infer user search goals for a query by clustering proposed feedback sessions. Feedback session is a series of both clicked and unclicked URLs and ends with the last URL that was clicked in a session from user click-through logs and proposed a novel optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs and cluster these pseudo-documents to get user search goals and depict them with some keywords. Finally this paper proposed a summarization technique to highlight the core information for the user to understand the information easily.

II. LITERATURE REVIEW

Huanhuan Cao and Daxin Jiang proposed a new method for suggesting queries to the user in a context-aware manner[1].Context-aware query suggestion approach has two steps.

1. Offline model learning step: Address data sparseness and queries are summarized into concepts by clustering a click-through bipartite. Concept sequence suffix tree is constructed from session data as the query suggestion model.
2. Online query suggestion step: User's search context is captured by mapping the query sequence submitted by the user to a sequence of concepts.

When a user submits a query q, context-aware approach captures the context of q which is represented by a short sequence of queries issued by the same user immediately before q. Then check the historical data and find what queries many users often ask after q in the same context. Those queries become the candidate suggestions. Then summarizing individual queries into concepts. Concept is a small set of queries that are similar to each other. To mine, concepts from queries, use URLs clicked for queries .This paper describe the mine concepts by clustering queries in a click-through bipartite. The click-through bipartite can help to find similar queries. Basic idea is that if two queries share many clicked URLs then they are similar to each other.

There are several challenges in clustering queries in a click-through bipartite.

1. A click-through bipartite from a search log can be huge.
2. The number of clusters are unknown. So another clustering algorithm should be needed to automatically determine the number of clusters.
3. Each distinct URL is treated as a dimension in a query vector. So data set is of high dimensionality.
4. The search logs increase dynamically. Therefore clustering needs to be maintained. No existing methods can address the above challenges simultaneously.

In clustering algoritham normalized centroid of cluster is determined from the queries.

M. Pasca and B.V Durme proposed a method for extracting relevant attributes from web[2]. This introduce a web query log rather than a web document. Different stratergies for extraction are class driven extraction and instance driven extraction. In class driven extraction the target class is specified by mentioning only the class name .In instance driven extraction the attributes of a given class can be derived by extracting and inspecting the attributes of individual instances from that class. The queries are used to extract knowledge base from Query logs. Extraction method used for the selection of candidate attributes, filtering of candidate attributes, ranking of candidate attributes. In selection of candidate attributes, a small set of linguistically motivated patterns extract potential pairs of a class label and an attribute from query logs. For each pair, weighted frequency is computed as a weighted sum of the frequencies of the input queries within the query logs. The frequencies of those matching queries are added to the counts computed initially for the pairs. Then converts the collected pairs into a smaller set of yet-unfiltered attributes that apply only to instances of the target classes. In filtering of candidate attributes a series of filters successively improve the quality of the sets of class attributes. First filter identifies and discards the attributes that are proper names or are part of longer proper names. Second filter discards candidate attributes that are deemed to be generic, in that they are simultaneously associated with many target classes. Third filter aims at reducing the number of attributes that are semantically close to one another within a class. In ranking of candidate attributes , weighted frequencies of the pairs of a class and an attribute that pass all filters are derived successively from the original frequencies in query logs. Computation takes the frequencies from query logs, weights based on which pattern they match, and then adds frequencies together as different entries are collapsed into identical pairs during pre-processing and then selection of attributes.

Chien Chin chen and Meng Chang Chen proposed topic anatomy system called TSCAN [3].This topic summarizes and associates the core part of topic. So the readers can understand the content easily. TSCAN extract themes, event and event summaries from topic documents.

Zheng Lu and Hongyuan Zha, Xiaokang Yang proposed a method for Inferring User Search Goals with Feedback Sessions [4].In this method feedback session is made from the user click through log and mapped to pseudo-documents.The user search goal is inferred by clustering these pseudo-documents. Performance of restructuring search results can be evaluated by evaluation criterion CAP. The evaluation result will be used to select the optimal number of user search goals.

Palvi Arora and Tarun Bhalla proposed a synonym based approach [5]. Today search is performed by searching the exact keywords entered by the user. But this may not result in the effective search because user may not know exact keywords. This paper introduces the following challenges:

1. It may not be able to perform effective search for queries which have no relevant synonyms in the synonym table.
2. It is difficult to implement synonym table containing separate entry for every keyword because there is no limit to keywords that a user can search for.
3. Large amount of time can be consumed in looking up synonym table if a keyword has many synonyms.

III. PROPOSED SYSTEM

The proposed method include mapping of feedback session into pseudo-point, clustering the pseudo-point and restructuring the web search result and finally apply summarization technique.

A. Feedback Sessions

A session for web search is a series of successive queries to satisfy a single information need and some clicked search results .Therefore, the single session containing only one query is introduced. Meanwhile, the feedback session in this paper is based on a single session. The proposed feedback session consists of both clicked and unclicked URLs and ends with the last URL that was clicked in a single session. Before the last click, all the URLs have been scanned and evaluated by users. Therefore besides the clicked URLs, unclicked ones before the last click should be a part of the user feedbacks. Inside the feedback session, the clicked URLs tell what users require and the unclicked URLs reflect what users do not care about. Unclicked URLs after the last clicked URL should not be included into the feedback sessions since it is not certain whether they were scanned or not.

B. Pseudo-Documents

Since feedback sessions vary a lot for different click-through and queries, it is unsuitable to directly use feedback sessions for inferring user search goals. Some representation method is required to represent the feedback sessions in a more efficient and coherent way. There may be many kinds of feature representations of feedback sessions. For example binary vector method to represent a feedback session. Different feedback sessions have different numbers of URLs. Therefore binary vectors of different feedback sessions may have different dimensions. Binary vector representation is not informative enough to tell the contents of user search goals. Therefore, it is not a proper methods. So a new method is needed to represent feedback sessions. The building of a pseudo-document includes two steps. They are described in the following section:

1) URLs representation in the feedback session:

In the first step, we first enrich the URLs with additional textual contents by extracting the titles and snippets of the returned URLs appearing in the feedback session. Each URL in the feedback session is represented by a small text paragraph that consists of its title and snippet. Some textual processes are implemented to those text paragraphs. Textual processes are transforming all the letters to lowercases, stemming and removing stop words. Each URL's title and snippet are represented by a Term Frequency-Inverse Document Frequency(TF-IDF) vector.

$$T_{ui} = \left[t_{\omega 1}, t_{\omega 2}, \ldots, t_{\omega n}\right]^T$$
$$S_{ui} = \left[s_{\omega 1}, s_{\omega 2}, \ldots, s_{\omega n}\right]^T$$

$T_{ui}$and $S_{ui}$are the TF-IDF vectors of URL's titles and snippets. ui the *i*th URL in the feedback session.ωj is *j*th term appearing in the enriched URL. The "term" is defined as a word or a number in the dictionary of document collections. t_{wj} and S_{wj} represent the TF-IDF value of the jth term in the URL's title and snippet. Enriched URL can be represented by the weighted sum of T_{ui} and S_{ui} namely feature representation of *i*th URL.

$$F_{ui} = \omega_t T_{ui} + \omega_s S_{ui} = [f_{\omega 1}, f_{\omega 2}, \cdots, f_{\omega n}]^T$$

Where ω_t and $\omega_s$are the weight of title and snippets.

2) *Forming pseudo-document:*

In order to obtain the feature representation of a feedback session, an optimization method to combine both clicked and unclicked URLs in the feedback session is proposed. Optimization can perform on each dimension independently as

$$f_{fs(\omega)} = \arg min_{f_{fs}(\omega)} \left\{ \sum_{N} [f_{fs(\omega)} - f_{ucm}(\omega)]^2 - \lambda \sum_{L} [f_{fs(\omega)} - f_{\overline{uci}}(\omega)]^2 \right\}, \quad f_{fs(\omega)} \in I_c$$

F_{fs} be the feature representation of a feedback session. $f_{fs(\omega)}$be the value for the term ω .

F_{ucm} (m=1,2,…M)and $F_{\overline{uci}}$ (l=1; 2, . . . ; L)be the feature representations of the clicked and unclicked URLs in this feedback session, respectively. Let $f_{ucm}(\omega)$ and $f_{\overline{uci}}(\omega)$be the values for the term ω in the vectors. We want to obtain such a F_{fs} that the sum of the distances between F_{fs} and each F_{ucm} is minimized and the sum of the distances between F_{fs} and each $F_{\overline{uci}}$ is maximized.

Let I_c be the interval $[\mu f_{uc(\omega)} - \sigma f_{uc(\omega)}, \mu f_{uc(\omega)} + \sigma f_{uc(\omega)},]$ and $I_{\bar{c}}$ be the interval $[\mu f_{\overline{uc}}(\omega) - \sigma f_{\overline{uc}(\omega)}, \mu f_{\overline{uc}(\omega)} + \sigma f_{\overline{uc}(\omega)},]$ where $\mu f_{uc(\omega)}$ and $\sigma f_{uc(\omega)}$, represent the mean and mean square error of $f_{uc(\omega)}$, respectively, and $\mu f_{\overline{uc}}(\omega)$ and $\sigma f_{\overline{uc}(\omega)}$, represent the mean and mean square error of $f_{\overline{uc}(\omega)}$, respectively. If $I_C \quad I_{\bar{C}}$ or $I_{\bar{C}} \quad I_C$ we consider that the user does not care about the term ω. In this situation, we set to be 0, as shown in $f_{fs(\omega)} = 0, I_C \quad I_{\bar{C}}$ or $I_{\bar{C}} \quad I_C$

λ is a parameter balancing the importance of clicked and unclicked URLs.

It is worth noting that people will also skip some URLs because they are too similar to the previous ones. In this situation, the "unclicked" URLs could wrongly reduce the weight of some terms in the pseudo-documents to some extent.

C. *Inferring user search goals*

With the proposed pseudo-documents,user search goals can be infered. This section describe how to infer user search goals and depict them with some meaningful keywords.

Each feedback session is represented by a pseudo-document and the feature representation of the pseudo-document is F_{fs} The similarity between two pseudo-documents is computed as the cosine score of $F_{fsi}$and F_{fsj}, as follows:

$$Sim_{i,j} = \cos(F_{fsi}, F_{fsj}) = \frac{F_{fsi} \cdot F_{fsj}}{|F_{fsi}||F_{fsj}|}$$

And the distance between two feedback sessions is

$$Dis_{i,j} = 1 - Sim_{i,j}$$

D. *Bisecting K-means clustering*

We cluster pseudo-documents by bisecting K-means clustering which is simple and effective. The bisecting K-means algorithm starts with a single cluster of all the documents and works in the following manner:

1. Pick a cluster to split.
2. Find 2 sub-clusters using the basic K-means algorithm.
3. Repeat step 2(bisecting step) for a fixed number of times and take the split that produces the clustering with the highest overall similarity. (Similarity is the average pairwise document similarity for each cluster, and we seek to minimize that sum over all clusters.)
4. Repeat steps 1, 2 and 3 until the desired number of clusters is reached. Each cluster can be considered as one user search goal. Since we do not know the exact number of user search goals for each query, we set K to be different values (i.e 1,2,..5) and perform clustering based on these values, respectively. The optimal value will be determined through the evaluation criteria.
5. Center point of cluster is computed as average of the all the pseudo document in the the cluster.

$$F_{centeri} = \frac{\sum_{k=1}^{ci} F_{fsk}}{C_i}, (F_{fsk} \in Cluster\ i)$$

where $F_{centeri}$ is the *i*th cluster's center and C_i is the number of the pseudo documents in the *i*th cluster. $F_{centeri}$ is utilized to conclude the search goal of the *i*th cluster.

6. Finally, the terms with the highest values in the center points are used as the keywords to depict user search goals.

E. Evaluation of web search result

Evaluation of user search goal inference is a problem, since user search goals are not predefined . The optimal number of clusters is still not determined when inferring user search goals. Therefore a feedback information is needed to finally determine the best cluster number. Therefore, it is necessary to develop a metric to evaluate the performance of user search goal inference objectively. If user search goals are inferred properly, search results can also be restructured properly. Restructuring web search results is one application of inferring user search goals. So, we propose an evaluation method based on restructuring web search results to evaluate whether user search goals are inferred or not. In this section, "Classified Average Precision" to evaluate the restructure results is proposed. Based on the proposed criterion, a method to select the best cluster number is also described.

1) Restructuring Web Search Results:

Since search engines always return millions of search results, it is necessary to organize them and make it easier for users to find out what they want. The restructuring web search results is an application of inferring user search goals. This paper describe how to restructure web search results by inferred user search goals at first and the evaluation based on restructuring web search results . Then, categorize each URL into a cluster centered by the inferred search goals. This paper proposed, categorization by choosing the smallest distance between the URL vector and user-search-goal vectors. In this way, the search results can be restructured according to the inferred user search goals.

2) Evaluation Criterion:

In order to apply the evaluation method to large-scale data, single sessions in user click-through logs are used to minimize manual work. From user click-through logs, get implicit relevance feedbacks, namely "clicked" means relevant and "unclicked" means irrelevant. Possible evaluation criterion is the average precision (AP) which evaluates according to user implicit feedbacks.

$$AP = \frac{1}{N+} \sum_{r=1}^{N} rel(r) \frac{Rr}{r}$$

where N+ is the number of relevant (or clicked) documents in the retrieved ones, r is the rank, N is the total number of retrieved documents, rel()s a binary function on the relevance of a given rank, and Rr is the number of relevant retrieved documents of rank r or less.

AP is not suitable for evaluating the restructured or clustered searching results.

The proposed new criterion for evaluating restructured results is describe Voted AP (VAP) which is the AP of the class including more clicks namely votes. If the numbers of the clicks in two classes are the same, then select the bigger AP as VAP. Assume that one user has only one search goal, then all the clicked URLs in a single session should belong to one class. A good restructuring of search results should have higher VAP. Still VAP is an unsatisfactory criterion.

Risk calculates the normalized number of clicked URL pairs that are not in the same class.

If the pair of the ith clicked URL and jth clicked URL are not categorized into one class then dij will be 1 otherwise 0.

$$Risk = \frac{\sum_{ij=1(i<j)}^{m} a_{ij}}{C_m^2}.$$

$C_m^2 = \frac{m(m-1)}{2}$ is the total number of the clicked URL pairs. Based on the above discussions, further extend VAP by introducing the above Risk and propose a new criterion "Classified AP," as shown below

$$CAP = VAP * (1 - Risk)^Y$$

CAP selects the AP of the class that user is interested in (i.e., with the most clicks/votes) and takes the risk of wrong classification into account. And Y is used to adjust the influence of Risk on CAP. Finally, utilize CAP to evaluate the performance of restructuring search results. Consider another case, if all the URLs in the search results are categorized into one class, Risk will always be the lowest namely 0; VAP could be very low. Categorizing search results into less clusters will induce smaller Risk and bigger VAP. More clusters will result in bigger Risk and smaller VAP. The proposed CAP depends on both of Risk and VAP.

F. Summarization of content

Text summarization method have been proposed to highlight the core information in the document. So that reader can easily understand the content. A topic anatomy system called TSCAN is used to extract the themes, event and event summaries from topic documents. Event is a disjoint subepisode of theme and they share similar context. Topic anatomy involves theme generation, event segmentation as well as summarization and evolution graph construction. In theme generation identify the theme of a topic from related documents. In event segmentation and summarization extract topic event and their summaries by analyzing the intension variation of themes over time. TSCAN organizes and summarizes the content of temporal topic described by a set of document. TSCAN model the document as symmetric block association matrix, in which each block is portion of document and treat each eign vector of matrix as theme embedded in the topic. Eign vectors are then examined and used to extract events and then summarizes from each theme. Temporal similarity(TS) function is applied to generate the event dependencies which are then used to construct the evolution graph of the topic.

IV. ARCHITECTURAL DESIGN

The framework of the method consists of three parts and it is shown in figure below. In the first part, all the feedback sessions of a query are first extracted from user click-through logs and mapped to pseudo-documents. Then, user search goals are inferred by clustering these pseudo-documents and depicted with some keywords. Since the exact number of user search goals is unknown in advance, several different values are tried and the optimal value will be determined by the feedback from the second part.

In the second part the original search results are restructured based on the user search goals inferred from the upper part. Then, evaluate the performance of restructuring search results by proposed evaluation criterion CAP. And the evaluation result will be used as the feedback to select the optimal number of user search goals in the first part.

In third part apply summarization technique to the search result. So it highlights the core information and represent the result in the form of document.

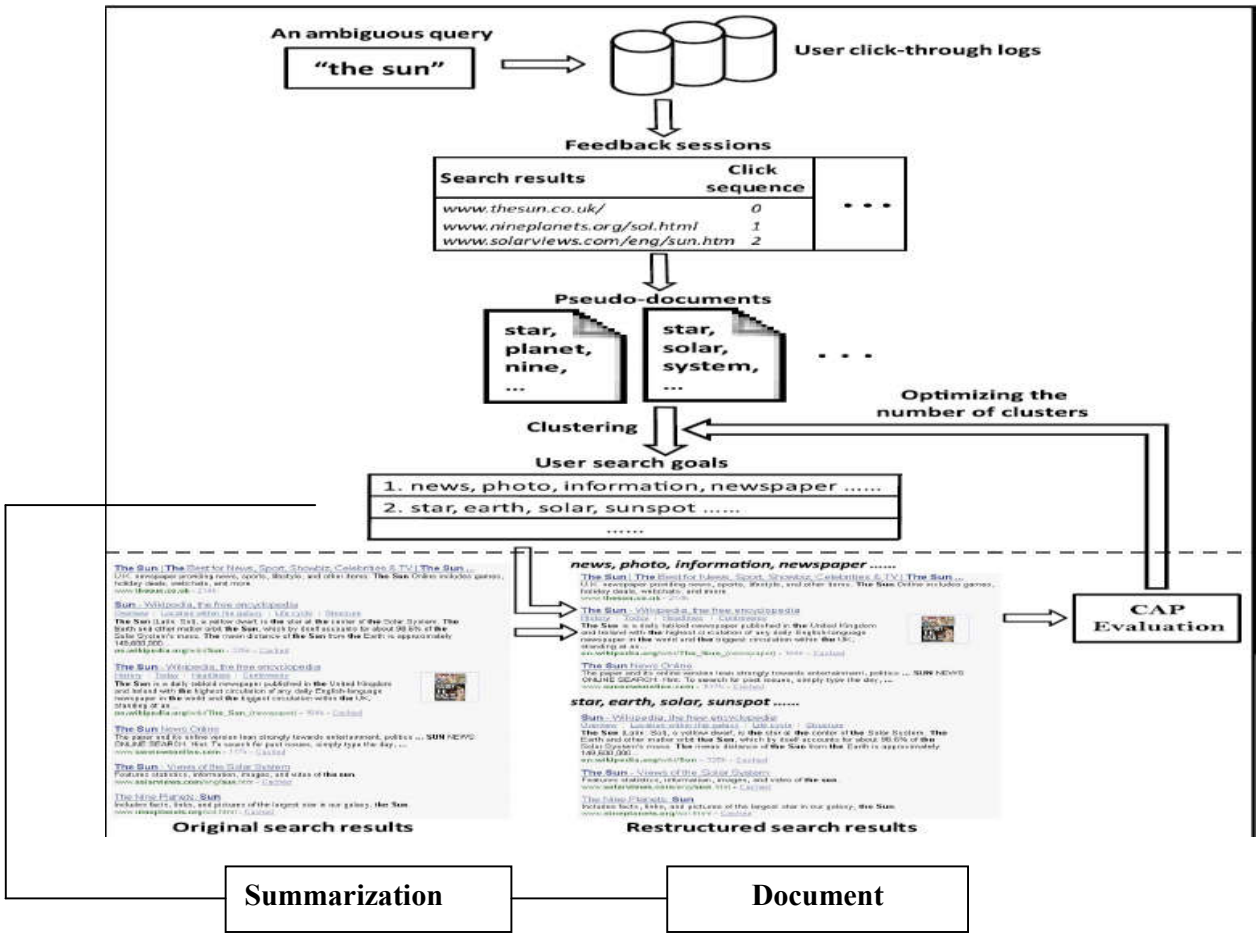

Fig 1. Architectural Design

V. IMPLEMENTATION DETAILS

The proposed method is implement in java. The software requirement include Eclipse, Tomcat server, WAMP server. Web forms are created in JSP.JSP is java server page. JSP creates in HTML, XML, CSS, java script, jquery, AJAX. WAMP server include phpMyAdmin, MySql, Apache ,php. phpMyAdmin is used for database creation. Queries are written in mysql. This method run in any operating system like Windows 7 (32 bit). Hardware Requirements include Processor (Above 1.5 GHz), Hard disc : 40GB,RAM : 1GB,Internet Connection.

VI. RESULT

This paper aim at discovering the number of diverse user search goals for a query. The registered user can search in to the system. The admin can control the user by accepting or rejecting them. Searched urls are saved in to the database. For calculating the feature vector, calculate the title vector and snippet vector. The weight of these vectors are multiplied with these vectors and find the sum of them to form the feature vector. This is used for pseudo-document creation. Then clustering is performed and then restructuring the result to get the search goal. Finally summarization is applied to get the result as document.

VII. CONCLUSION

In web search applications, queries are submitted to search engines to represent the information needs of users. In this paper, a novel approach has been proposed to infer user search goals for a query by clustering its feedback sessions represented by pseudo-documents.

Feedback session is created by using both clicked and unclicked URLs. The feedback session is mapped into pseudo document. User search goal can be discovered by clustering these pseudo documents . Finally summarization technique is applied to obtain result and to highlight the core information in the document. This reduces the complexity. The running time depends upon the feedback session.

REFERENCES

[1] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "Context-Aware Query Suggestion by Mining Click-Through," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '08), pp. 875-883, 2008.

[2] M. Pasca and B.-V Durme, "What You Seek Is what You Get:Extraction of Class Attributes from Query Logs," Proc. 20th Int'l Joint Conf. Artificial Intelligence (IJCAI '07), pp. 2832-2837, 2007.

[3] Chien Chin chen and Meng Chang Chen "TSCAN:a content anatomy approach to temporal topic summarization"

[4] Zheng Lu ,Hongyuan Zha, Xiaokang Yang, "A New Algorithm for Inferring User SearchGoals with Feedback Sessions".

[5] Palvi Arora and Tarun Bhalla " A Synonym Based Approach of Data Mining in Search Engine Optimization". [6] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '00), pp. 407-416, 2000.

[7] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "Varying Approaches to Topical Web Query Classification," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development (SIGIR '07), pp. 783-784, 2007.

[8] H. Chen and S. Dumais, " Bringing Order to the Web: Automatically Categorizing Search Results, " Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 145-152, 2000