SEMANTIC ANALYSIS OF QUERY RESULTS FOR WEB INFORMATION RETRIEVAL

NIMISHA A. MODI

Department Of Computer Science, VNSGU, Surat.

nimishamehta@yahoo.com

ABSTRACT: *Search Engines are maintaining the index for the enormous collection information available in web pages. The meta-search engines or personalized search agents are introduced to utilize the valuable computing resources for optimizing various aspects of web IR rather than just repeating the efforts for creating and indexing their own repository. These search utilities generally include pre-processing techniques for refining user query or post-query processing analysis like clustering and linguistic analysis to optimize search precision. The paper focuses on the personalized search agent that maintains user profile and identifies the semantic context of user's requirement based on her profile. The existing methods for web usage mining use this user's profile before finding similarity of query phrase and documents and based on this semantic analysis the query is being expanded. The paper highlights the various query expansion techniques and suggests an approach to utilize the user profile at post-query extraction stage rather than pre-query extraction stage.*

KEYWORDS: *Web Search Agent, Semantic Context, Query Expansion, Web Usage Mining.*

1. INTRODUCTION

Web Search Engines [1] are constantly collecting enormous amount of information from web pages using crawlers. They consume huge amount of recourses and time to maintain the updated index for this information.

Meta-search engines [2] are web search utilities that do not create and maintain their own repositories for huge amount of web pages. When the user submits the query, meta-search engine transmits user query into several individual search engines simultaneously and collects result from all. These results are then compiled to provide optimal response to user. Meta-search optimization technology generally includes pre-processing and post-query processing analysis that includes link structure analysis [3] and semantic analysis. They utilize their valuable resources in attempt to dig deeply within the initial results of base search engines. Generally they apply some fancy textual analysis and display results in more user friendly way.

The semantic analysis process is examined in context of personalized search agent. Personalized search agent is a supporting system for the web information seekers that works as the interface between users and search engine, and assists the web users to intelligently retrieve information from the web. The search agent makes use of link structure analysis and semantic analysis techniques on the results that are returned by search engines. The paper evaluates the existing techniques for analysis of semantic context. Existing techniques are basically pre-processing techniques and applied before similarity between documents and query phase is found. The paper introduces the concept of using this semantic analysis at post query processing stage i.e. after getting the initial results on the given query.

2. SEMANTIC CONTEXT FOR WEB IR

Conventional technique for Information Retrieval (IR) is to lexically match the query terms with terms in target documents. The pure lexical matching may be inefficient for the information that represented via a huge variety of terminology. The presence of polysemy terms may return the irrelevant document to user which pertaining to totally different perspective and thus decreases the search precision. The relevant document, consisting of the required information but using different vocabulary (or synonyms) than the query term, are failing to match the query phase and as a consequences the recall rate also decreases.

For the global collection of enormous data on web where at one side vocabulary is varying in huge amount, while on other side majority of web user submits very short query that rarely specifies the precise meaning of her information need. In addition to the effect of polysemy and synonyms, keyword spamming also misguides search engines. Specifically in case of broad topic queries for which huge number resources are available on web, but user is interested in only few but most relative and important results among them. The challenge is raised amongst the web IR utilities to obtain user's satisfaction in minimum explicit interaction. This

challenge leads to exploitation of various techniques to match the semantic context of user query and document collection. The techniques that explore the semantic similarity between the query and document include statistical analysis, use of thesauri resources, local analysis using feedback and usage log analysis.

Statistical Technique

Latent semantic indexing (LSI) is an indexing and retrieval method that uses a mathematical technique to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. LSI [4] is based on the principle that words that are used in the same contexts tend to have similar meanings. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts. Such techniques help to find the matching document even if they do not explicitly contain the noun phase of query.

Thesaurus

Thesaurus provides the synonyms and semantically related words and phases. The thesaurus or dictionaries are generally created with human defined knowledge; the knowledge can be utilized in order to uncover indirect connections between query terms and document content. Thesaurus [5] can be easily used for search utilities with purpose of including synonyms or statistically related terms to increase recall rate, but some time at the cost of precision as ambiguous term may introduce irrelevant statistically correlated terms.

Local Analysis

Local analysis marks higher correlation between the queries and selected (here clicked) document in response to that query. Two variation of local analysis are Relevance Feedback and Pseudo Relevance Feedback [6]. Relevance feedback is cyclic process, which in first run requires the user's feedback on the relevancy of documents from the initial result set. This feedback is either explicitly provided by user using a binary or graded relevance system or implicitly inferred from user behavior such as whether or not they select document for viewing, the duration of time spent viewing a document etc… The query phase is then refined based on selected documents considering them as relevant based on the hypothesis - if a set of documents is often selected for the same queries, then the terms in these documents are strongly related to the terms of the queries. So, in the next run the retrieval process becomes more specific towards context of the initially selected result.

In absence of sufficient feedback, one can go for pseudo relevance feedback that blindly considering that the documents returned by base search engine are relevant. Pseudo Relevance Feedback or Blind Relevance Feedback is iterative process that uses relevance feedback technique without explicit user input by blindly assuming the top n retrieved documents as relevant. To reformulate the query, it includes those terms from documents that are correlated with that query terms.

Web Usage Mining

The features of IR utilities on web introduces the new dimension to local analysis i.e. analysis of query log using the techniques from web usage mining [7] that is based on the users interaction i.e. query submission and documents selection in response to that query. This information can be tracked in query log for search engine.

Hang sui [8] proposed an effective method for context analysis based on user's interaction on web that is recorded in the web query logs. The web search agent can collect information about clicked (relevant) document form user log, which represents user interactions with searching systems via web usage mining.

Web usage mining refers to the discovery of user access patterns from web usage logs. Web servers record and accumulate data about user interactions whenever requests for resources are received. The web access logs are analyzed to understand the user behaviors and used for improving the design of this huge collection of resources. Three main areas in web usage mining driven by the applications of the discoveries are – General-Access Pattern Tracking, Customized Usage Tracking and User Behavior Tracking.

General-Access Pattern Tracking analyzes the web logs to understand access patterns and trends, while Customized Usage Tracking analyzes individual trends. Its purpose is to customize web sites to users. All of us have some experiences about the consequences of general access pattern tracking as well as customized usage tracking while searching and ordering books at amazon.com.

Web usage mining is implemented by various web IR tools to personalize the search results using user behavior tracking. User Behavior Tracking analyzes web access data in general for individual users.

3. USER BEHAVIOR TACKING FOR SEMANTIC ANALYSIS

Basically user actions on web are being tracked to identify user's interest, to make prediction about her subsequent requests and to personalize the search results. Information retrieval system uses the machine learning algorithms to develop the usage profile that accumulate learning from users' past search activities with user behavior tracking.

The agent system considers the user profile as a set of terms that is previously used by individual user. This user profile is used to reveal the semantic of user's interactions and thus personalize the web search results. User profile that reflects the user's interest or query semantic could be used at different stages of information retrieval process.

The usage data can be used either at pre-processing stage or post-processing stage to filter queries or initial results for presenting the most relevant resources to the users in the response to their queries.

Pre-processing Stage: Query Expansion

Existing research focuses on the use of usage profile at pre-processing stage that applies before the similarities between the query and the documents are measured. Hang [8] proposed a method for query expansion by mining uses' web access logs. Query expansion techniques are integrated with information retrieval systems to make the user's query more specific towards user's perspective of her information need with the objective of identifying the semantic context and receiving well précised results. Hang proposed this query expansion by analysis of usage log. As the short query provided by user does not carry sufficient information to convey her information need unambiguously, the user profile is being mined and based on that the query is being expanded.

As for example, while searching for the query 'java', web IR tool decides the semantic of the user's requirement for 'java' using the usage profile of the user. If usage profile is specific to developer or student that is generally using terms related to programming languages or computer, the corresponding terms are added to query. So the query is expanded to either 'java program' or 'java language' or 'java programming tutorial' or any other combination and then it will be submitted to query engine. On other side, if the user is travel agent or frequently searching for 'tourist attractions' or 'islands', the query may intended for 'java island'.

Basically usage analysis is integrated with information retrieval at pre-query processing stage. As, shown in figure 1, the semantic analysis is done for expanding the query and after that the expanded query is submitted to the query engine of IR system for finding its relevance with documents in corpus.

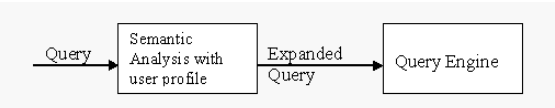

Fig. 1 Semantic Analyses for Query Expansion

Post-processing Stage: Search Result Mining

The personalized search agent employs a novel approach that integrates this semantic analysis with web IR after the initial results are collected and before presenting these results to the user as shown in figure 2.

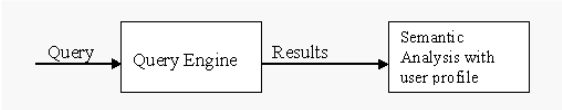

Fig. 2 Semantic Analyses for Post-Processing the Results

The personalized search agent uses these web usage data at post-processing stage. Post-processing techniques work on the initial results that are obtained after calculating the similarities between the query and the documents. The initial results are dug deeply for matching semantic context.

4. QUERY EXPANSION VS SEARCH RESULT MINING

Query expansion is most wildly used for exploring semantic context in IR. The limitation of query expansion or refinement is - the search becomes narrow (specific) search. Possibly, the user wants to search for some other perspective of the topic. For example, a person in biology may require a search for the 'virus' that had infected her computer system. In such cases, the query expansion approach needlessly limits the search results to biological context and fails to satisfy the user's requirement. The proposal for the post-processing approach targets two objectives - grouping of Search Results and Search result caching

Grouping of Search Results

The core effort is for grouping of the search results. Here, it makes broad context search before performing any semantic analysis, as user's context of the current search activity may or may not be according to her previous search activity. User may require searching for some other perspective of topic i.e. a researcher in biology finds for the computer 'virus'. So, the post-processing approach is to collect the results in response to broad query rather than on the expanded (specific) one, and further re-organize the results in different group (a set of documents) according to their semantic similarities.

Search Result Caching

Second objective is to propose a model that can be used by general purpose search engine. Evangelos [9] had traced a popular Search Engine (Excite) to show that there is a significant locality in the queries. More than 20-30% of the queries have been previously submitted by the same or a different user. So defining query result caching and pre-fetching policies becomes another major issue for web search engine.

Search engines make a cache for query results for frequent queries. With pre-processing approach of query expansion, the search utility can cache the raw query result for expanded query which will only

beneficial if that topic is searched in the same context again. While with post processing approach, in place of raw query results one can store the results in various semantic groups found after post-processing the raw results of frequent and broad topics query. When user's search on such query, search engine have groups of web pages on various semantic of that topic in cache, so it just needs to map the proper group of results to user's context and present them amongst the user.

5. CONCLUSION

The personalized search agent uses the post-processing approach of search result mining to achieve above objects. It represents various semantic of the query in different groups to user. The usage profile is not used for narrowing the query, rather it this profile is used to organize the semantic of user's routine as top among result set. As per the previous example, when a person in biology submits the query 'virus', the proposed approach does not limit the search results to biological context prior to submitting the query. It makes a broad (general) search on the term 'virus'. Semantic analysis is performed to arrange the results according to relevance with user's interest. Thus, semantic analysis gives the preferences to pages related to biological virus. The highly authoritative pages about computer virus are not excluded from the results, but the sequence of their appearance comes after the pages related to biological virus.

6. REFERENCES

[1] Sergey Brin, Lawrence Page, (1998) 'The anatomy of a large-scale hyper-textual Web search engine', *Computer Networks and ISDN Systems*, 33:107-117.

[2] Sandra Zabala, Gabor Loerincs, Yubelsi Bello, Victor Dias, (2001), 'CALVIN: A Personalized Web-Search Agent based on Monitoring User Actions', *Web Databases Workshop,* 2001, (http://doesen8.informatik.uni-leipzig.de/webdb/wien/papers.html).

[3] Nimisha Modi, Dr. A.A. Desai, (2008) 'Analyzing Spatial Locality on Web Graph: Concept Searching with Search Agent' *proceeding of International Conference on Emerging Technologies and Applications in Engineering, Technology and Sciences*, 13-14 January 2008.

[4] Barbara Rosario (2000) 'Latent Semantic Indexing: An overview' *Final Paper, INFOSYS 240 Spring* 2000.

[5] S. Liu, F. Liu, C. Yu, W. Meng. (2004) 'An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases'. *ACM SIGIR Conference, pp.266-272, Sheffield, UK, July 2004*

[6] Yuanhua Lv, ChengXiang Zhai, Wan Chen, (2011) 'A boosting approach to improving pseudo-relevance feedback' SIGIR 2011: 165-174

[7] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan (2000) 'Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data',*SIGKDD Explorations*.

[8] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, Wei-Ying Ma (2003), 'Query Expansion by Mining User Logs', *IEEE Transactions on Knowledge and Data Engineering, July/Aug 2003, Vol. 15, no. 4, pages 829-839*.