J. Appl. Environ. Biol. Sci., 4(7S)271-276, 2014

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

A Survey On Diversification Techniques For Unabmiguous But Under-Specified Queries

Muhammad Shoaib Farooq[1,2], Sher Afzal Khan[2], Farooq Ahmad[3], Kamran Abid[4], Uzma Farooq[1], *Adnan Abid[1]

[1]Department of Computer Science, University of Management and Technology, Lahore, Pakistan.
[2]Department of Computer Science, Abdul Wali Khan University, Mardan, Pakistan.
[3]Faculty of Information Technology, University of Central Punjab Lahore, Pakistan.

Received: September 1, 2014
Accepted: November 13, 2014

ABSTRACT

The amount of data placed on the web has been greater than before and is increasing rapidly day by day. Web searching, the huge size of result set, ranking and presentation of results becomes important. Mostly users only look at the first page of available results and neglect the rest. To improve user's satisfaction, the listed results should be relevant to the search topic and different from each other. Web search effectiveness and user satisfaction can be improved by providing various results of the search query in a certain order of relevance and concern. The purpose of diversification is to avoid presenting similar results and introducing diversity with variety of search results. In short, diversification re-ranks the relevant search. Diversification and personalization methods are common approach to deal with the one-size-fits-all model of web search engines. In this paper, we discuss different techniques and algorithms to deal with underspecified query.
KEYWORDS: Web Search, relevance, diversity, personalization.

1. INTRODUCTION

These queries are unambiguous in the sense that the sense of these queries is clear; there is only one way to read or understand these queries. They refer to an unambiguous entity however, it is not clearly specified what the user wants to know about the entity. Consider, for example, the query "Madonna". In Fig. 1 there is no ambiguity in what the query means but still it is not clear what the user wants to know about Madonna does he want to watch the music videos, read news, find song lyrics, or purchase songs at the iTunes store? The user's intent is not stated. For such queries, the search engine needs to focus on determining the underlying intents behind the underspecified query and create a result list to cover these different intents accordingly.

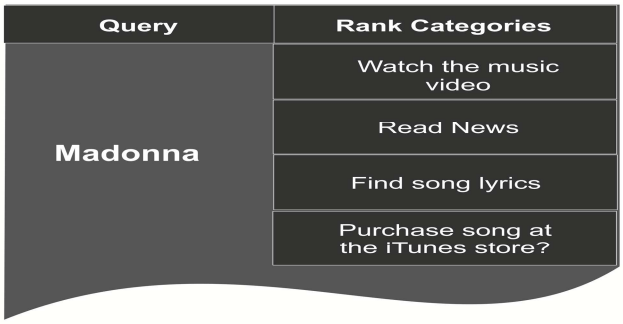

Figure 1 Result Set of the Madonna Query

In our study we found that the problem of unambiguous but underspecified queries has been addressed using two different techniques.

* **Corresponding Author:** Adnan Abid, Department of Computer Science, University of Management and Technology, Lahore, Pakistan.

1. **Personalized diversification**: This procedure is built on two steps, first the system requirement gathers personal information from user profile, and secondly the diversification must be applied on the result set which is saved in first step [1],[5]. Such type of technique is discussed in section 2
2. **Query Log**: In Query log scheme the system automatically suggests a set of queries, based on the original query; the proposed suggestion represents different possible interpretation [2], [4]. Such kind of procedure is discussed in section 3.

2. Personalized Diversification

Personalization is the process of presenting the right information to the right user at the right moment. In order to learn about a user, system must collect personal information, analyze it, and store the results of the analysis in a user profile. Commercial systems tend to focus on personalized search using an explicitly defined profile, for example, users are asked to select the categories of topics which they are interested in and the search engine applies this information during the retrieval process [3].

User profiles can also be divided in two other groups

1. **User's preferences** (e.g., search engines preferred, types of documents)
2. **User's interests** (e.g., sports, photography.).

The basic principle behind personalized search is simple. When a user goes to WSE and type in a search query, WSE stores the data. As user returns to the engine, a profile of user's search habits is built up over time. With this information, WSE can understand more about his interests and serve up more relevant search results.

For instance, let's say that user has shown an interest in the topic of sport fishing in its search queries, while his neighbor has shown an interest in musical instruments in his search queries. Over time, as these preferences are made clear to the engine, the user's personalized search results for the term 'bass' will largely be comprised of results that cover the fish while his neighbor's results for 'bass' will be comprised of results that primarily cover the musical instrument.

At present, the user need to have signed up for a WSE service for the results to be personalized. Such services include Gmail, Ad Words, Google Toolbar, and many others. By default, as long as user is signed in to one of these programs, his personal search data will be collected, WSE already places a cookie, or unique identifier, on the machine of anyone who types in a search query on it would not be hard for them to use that information, rather than the WSE account, to collect individual user data and personalize results.

2.1 User Profile

A user profile is constructed from Web pages browsed by the user. However, this technique focuses on using the user's search history.

Profile Based on User's Preferences. User profile, based on user's preference, runs as a background process on the user's machine. The application can retrieve results immediately after a query has been submitted. The agent running in the background should help users to reduce the amount of time spent on a search. In this case, the profile is supplied to an agent that can automatically gather information on behalf of the user.

Profiles Based on User's Interests: Interests-based profiles are more determined than preferences-based profile, because they try to extract from documents topics and subjects that match user's needs. User browsing histories are the most frequently used source of information to create interest profiles. In this system, user profiles are implicitly created based on browsing histories rather than explicitly created from user input. The hierarchy of user's interests is created using a clustering approach. The set of interests represents a user profile which can be used to automatically search for information, for filtering or to personalize the way of showing information. The analysis is based on documents collected from web pages visited and emails received or sent [3].

Classical Web Search Model. In Figure 2 the user pass the simple Query "Queen" which is unambiguous but underspecified query by using classical web search model, the result set of the query is based upon query log technique.

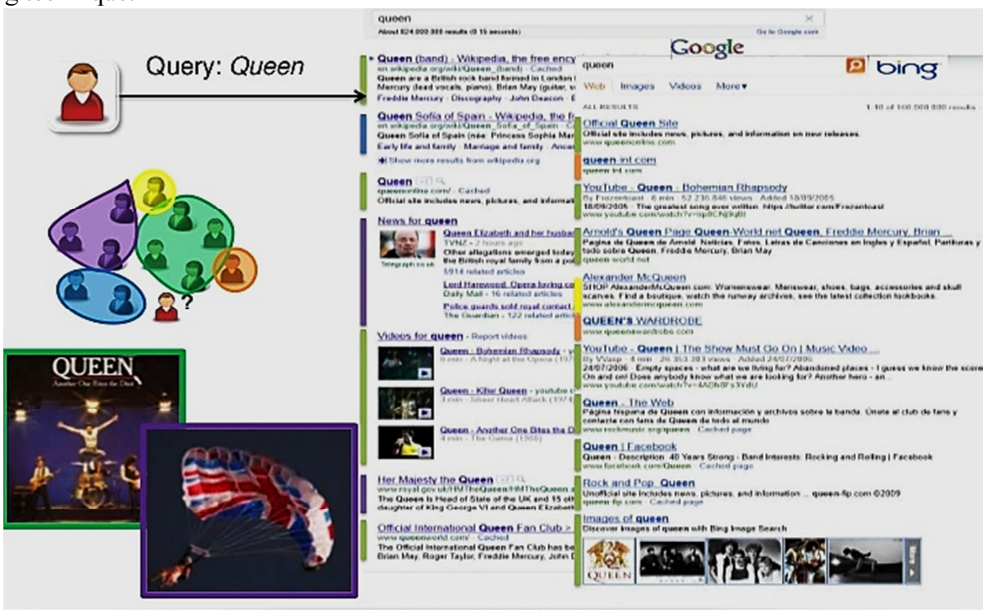

Figure: 2 Result sets of personalized web search model

Figure 2 shows the result sets of two famous web search engines Google and Bing, In Google log four user's used the web site Queen (Band), three user's used "Queen Sofia of Spain" and two user's used "Queen official site", In Bing log four user used the web site "official queen site", three users used "queen-int com" and two users used "You-tube Queen".

The ordering of results set are based upon the no of user's interested in the past on that particular query.

Personalized Web Search Model. Figure 3 shows that by using personalized web search model, query goes to user profile and from where it will map to user's interest and at the end personalized ordering will have to be done. In this example query "Queen" goes to user profile, from there using personalized ordering result will come according to Figure 3.

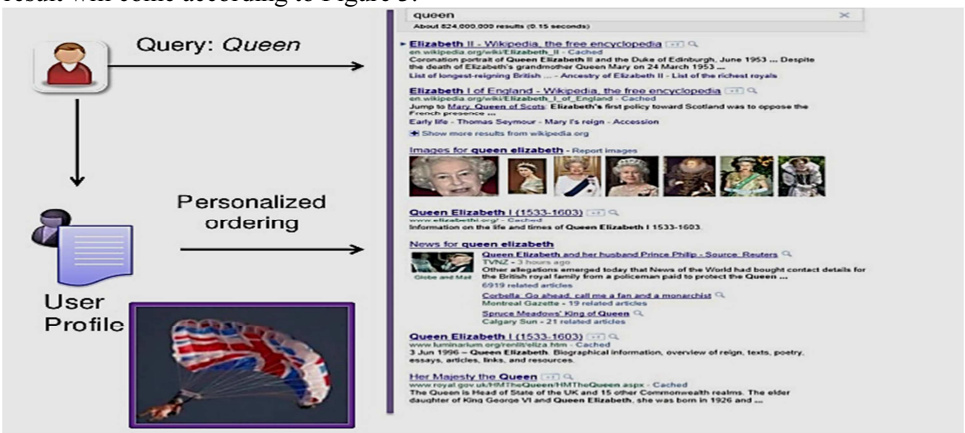

Figure: 3 Result sets of personalized web search model

Diversified Web Search Model. Figure 4 shows the result set of query "Queen" by using diversification model. One positive aspect of this model is that all links are relevant and almost different from each other.

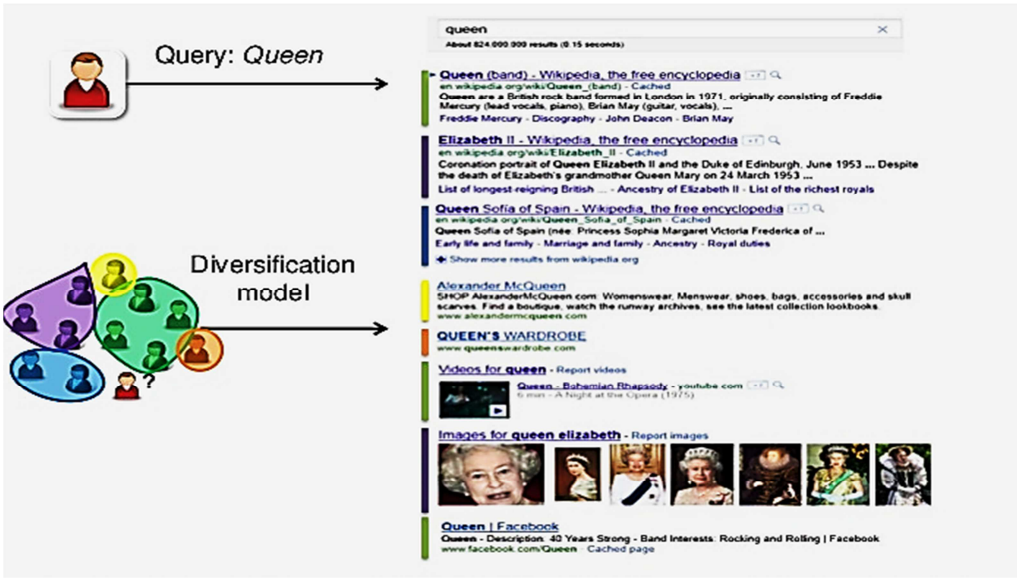

Figure: 4 Result set of diversified web search model.

2.3 Diversify personalization Framework:

i. Probabilistic model:

$\boldsymbol{p(c|q)}$: Relation between category and the query (e.g. popularity of certain aspect in a query)

$\boldsymbol{p(q|d), p(d|q)}$:Relation between document and query (e.g., ranking score of document)

$\boldsymbol{p(c|q), p(d|c)}$**:** Relation between document and category (e.g., document classification)

ii. IA-Select:

This framework use the following equation (1)

$$fs(d) = \sum_c p(q|d)p(c|d)p(c|q)\prod_{d'\in S}(1 - p(q|d')p(c|d')) \quad (1)$$

Where

$p(q|d)p(c|d) =$ Document relevance

$p(c|q)\prod_{d'\in S}\left(1 - p(q|d')p(c|d')\right)$= Novelty

iii. Personalized IA-Select:

Adding a user component in Equation (1)

$$fs(d) = \sum_c p(q|d,u)p(c|d,u)p(c|q,u)\prod_{d'\in S}(1 - p(q|d',u)p(c|d',u)) \quad (2)$$

iv. xQuad:

This framework use the following equation (3)

$$fs(d) = (1-\lambda)p(d|q) + \lambda\sum_c p(c|q)p(d|c)\prod_{d'\epsilon S}\left(1 - p(d'|c)\right) \quad (3)$$

Where

$p(d|q)$ = document query relevance

$\sum_c p(c|q)p(d|c)$= document topic relevance

$\prod_{d'\epsilon S}\left(1 - p(d'|c)\right)$ = Novelty

Allowing adjusting degree of diversification (λ)

v. Personalized xQuAD:

Adding a user component (4)

$$fs(d,u) = (1-\lambda)p(d|q,u) + \lambda \sum_c p(c|q,u)p(d|c,u) \prod_{d' \epsilon S} \left(1 - p(d'|c,u)\right) \quad (4)$$

3. Query Log

To implement personalization the search system records and analyses some additional information concerning the users together during helping the user operations earlier to submitting the query. This information can have many forms such as: user profiles, search history or click history of users. By using personalization following issue arise:

- It may be difficult or impossible to collect information or data from the user to effectively build their profile
- Gathering such data usually violates user privacy

Another approach to deal with under specification of user information need is query suggestion and this approach is based on query log i.e. the system automatically suggests a set of queries, based on the original query, such the each proposed suggestion represents different possible interpretation [2].

3.1 Algorithm based on Time Succession:

This framework [2] is used for automatic query suggestion based on historical query logs. A graph $G(V,E)$ is built so that each vertex $v \in V$ represents some unique query found in logs and there is an arc $(v,w) \in E$ if the following conditions are satisfied:

- The query w occurs at least once directly after the query v in the logs and was submitting by the same user
- The difference in time between submitting the queries is not higher than some threshold T , that is one of the parameters

3.2 Diversification Framework:

To improve the coverage of different interpretations or aspects of an underspecified query in the set of query suggestion, MMR ("Maximal Marginal Relevance") algorithm, usually used to diversify recommendations. By using the original formulation of MMR, This framework proposes to adopt this method to problem of diversified query suggestion.

Thus the adopted formula is as follows:

$$qSuggMMR = argmax_{qi \in R \setminus S}[\lambda sim_1(q_i,q) - (1-\lambda) max_{qj \in S} sim_2(q_i,q_j)] \quad (5)$$

Where q is the original (underspecified) user query and R is the set of candidate query suggestions and qi represent candidates to be selected to the final set of query suggestions S.query suggestion method described in section 3.1 to compute the initial set of candidates R.

4. DISCUSSION

In our study we found that the problem of unambiguous but underspecified queries has been addressed using two different techniques.

First technique is suggested personalized diversification. In this technique, a user profile is constructed from Web pages browsed by the user. User profile, based on user's preference runs as a background process on the user's machine. User's query match from the user's preference which are mentions in user's profile and based on this method relevant result set will come. This framework used two state of art IA-Select and xQuAD diversification algorithms. The personalized forms of IA-Select and xQuAD offer a framework that supports the combination between diversity and personalization.

David Vallet performs different experiments, a general summary of the results show that the best performing method in term of diversity metrics is PxQuAD. The fact will change with respect to the baseline both in terms of topic and user relevance. In subtopic recall the PIA-Select is the best performing one.

Second technique is suggested query log. This framework applied and combined the algorithm for query suggestion and diversification. This technique was diversified with the MMR-based diversification algorithm. MarcineSydow performs different experiments, a general summary of the results show that MMR perform very fine in the absence of user personal profile data. He also associates non-diversified results with MMR-based algorithm. MMR -based algorithm improved the level of diversification.

5. Conclusion And Future Work

This Framework presented two novel approaches to improve user experience in search engines by means of automatic, log-based query suggestion and User profile, in particular, diversified query suggestion. There are two state of art diversification algorithms xQuAD and IA-Select which are used for personalized diversification. MMR is used for query log method. There is room to find better categorization in personalized diversification in future.

REFREENCES

[1] David Vallet and Poblo Castells: Personalized Diversification of Search Results, SIGIR'2012.

[2] MarcineSydow and Krzysztof Ciesielski, Institute of Computer Science, Polish Academy of Science, Warsaw, Poland: Introducing Diversity to Log-Based Query Suggestions to Deal with Underspecified User Queries.SIIS 2011.

[3] Mirco Speretta , Udine University Udine, Italy : Personalized Search Based on user search. 2000

[4] Fillip Radlinski, Control University Ithaca,NY,USA and Susan Dumais, Microsoft Research Redmond, WA,USA: Improving Personalized Web Search using Result Diversification.SIGIR '2006.

[5] David Vallet, Pablo Castells:On Diversifying and Personalizing Web Search .SGIR'2011.