New Assessment Criteria for Query Suggestion

Zhongrui Ma  
Renmin University of China  
zhongrui.ma@gmail.com

Yu Chen  
Microsoft Research Asia  
yu.chen@microsoft.com

Ruihua Song  
Microsoft Research Asia  
rsong@microsoft.com

Tetsuya Sakai  
Microsoft Research Asia  
tesakai@microsoft.com

Jiaheng Lu  
Renmin University of China  
jiahenglu@ruc.edu.cn

Ji-Rong Wen  
Microsoft Research Asia  
jrwen@microsoft.com

ABSTRACT

Query suggestion is a useful tool to help users express their information needs by supplying alternative queries. When evaluating the effectiveness of query suggestion algorithms, many previous studies focus on measuring whether a suggestion query is relevant or not to the input query. This assessment criterion is too simple to describe users’ requirements. In this paper, we introduce two scenarios of query suggestion. The first scenario represents cases where the search result of the input query is unsatisfactory. The second scenario represents cases where the search result is satisfactory but the user may be looking for alternative solutions. Based on the two scenarios, we propose two assessment criteria. Our labeling results indicate that the new assessment criteria provide finer distinctions among query suggestions than the traditional relevance-based criterion.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: query formulation

General Terms
Measurement

Keywords
Assessment criteria, query suggestion

1. INTRODUCTION

Query suggestion is widely adopted in commercial search engines to supply alternative queries and assist users to achieve their search goals. When evaluating the effectiveness of query suggestion methods, some previous works \[1\,2\,6\] assign relevant or irrelevant labels to the suggestion queries, and then compute metrics such as precision, recall, MAP.

This evaluation method, however, is too rudimentary to capture users’ requirements precisely, and various refinements have been proposed in the literature. For example, Jain et al. \[5\] require a good suggestion to be not only relevant but also useful. The suggestions return zero search results or are synonyms of the input query are treated as bad. Bhatia et al. \[3\] also regard suggestions that are almost duplicate to the input query as irrelevant. Yang et al. \[7\] adopt a five-point score to distinguish suggestions from “Precisely related” to “Clearly unrelated”. Baraglia et al. \[2\] define a suggestion as useful if it “can interpret the possible intent of the user better than the original query”.

We argue that we should emphasize the usefulness instead of relevance because a relevant suggestion is not necessarily useful to the user. Therefore, we attempt to describe the usefulness of a suggestion by our proposed assessment criteria.

2. THE ASSESSMENT CRITERIA

Whether a suggestion is useful or not depends on the situation that the user encounters in her/his search session. We believe there exists at least two different scenarios:

Scenario 1 The search result of the input query is unsatisfactory. In this case, the user expects a suggestion that can return better search results to meet his/her information need. For example, for an input query \( q = \text{“aol instant mess”} \), two suggestion queries \( q_1 = \text{“aol instant messenger”} \) and \( q_2 = \text{“aol aim”} \) (here, “aim” is the acronym of “AOL Instant Messenger”) are both labeled as relevant to \( q \) by the assessor. When we submit these three queries to a search engine and check their results, we found that \( q_1 \)’s result matches \( q \)’s intention better than the results of both \( q \) and \( q_2 \). Therefore, \( q_1 \) is more useful than \( q_2 \) in terms of retrieving relevant documents.

Scenario 2 The search result of the input query already contains desirable documents. In this case, the user expects a suggestion that is related to the input query, but has some difference in topics, in order to broaden his/her knowledge. For example, when the input query is “aol instant messenger”, whose search result is good enough, returning “windows live messenger” and “yahoo! messenger” as alternative solutions might be more useful than returning “aol aim”.

According to the two scenarios, we propose two judgment guidelines for human assessors.

2.1 Utility Judgment

The utility judgment address scenario 1 by measuring whether a suggestion query \( q_s \) can retrieve more relevant web documents than the input query \( q \), based on the information need described by \( q \). To decide the utility judgment of \( q_s \) for \( q \), an assessor must compare the search results of both \( q \) and \( q_s \). Then she can put \( q_s \) into one of the following three categories, based on the comparison:

- **Better**: \( q_s \) retrieves more relevant documents than \( q \);
- **Worse**: \( q_s \) retrieves fewer relevant documents than \( q \);
- **Same**: The quality of \( q_s \)’s search result is similar to that of \( q \).

Although some of the existing works, e.g. \[3\], also require assessors to read the search results of \( q \) and \( q_s \), the search results are
only for understanding the intention of \( q \) and \( q_i \) in order to decide whether they are relevant. To the best of our knowledge, none of them compare the quality of the two search results.

### 2.2 Relationship Judgment

The desirable suggestions in scenario 2 should be related to the input query and have some difference in topic. Instead of using this vague expression directly, we ask assessors to classify the relationship between \( q \) and \( q_i \) into one of the following five categories:

- **Same intention**: If \( q_i \) has the same search goal with \( q \). E.g., \( q = \text{"msn messenger"} \) (also for all the following examples), and \( q_i = \text{"windows live messenger"} \);
- **Generalization**: If \( q_i \) is more general than \( q \). E.g., \( q_i = \text{"instant messenger"} \);
- **Specialization**: If \( q_i \) is more special than \( q \). E.g., \( q_i = \text{"msn messenger download"} \);
- **Peer**: If \( q \) and \( q_i \) are about two different things, but belong to a common topic. E.g., \( q_i = \text{"yahoo! messenger"} \). Both \( q \) and \( q_i \) belong to the class of instant messenger software;
- **No association**: \( q \) and \( q_i \) are irrelevant to each other.

Suggestions in generalization, specialization, and peer are the desirable ones, because they are related to the input queries and have some differences in topic. Suggestions in same intention are less useful as they have no difference in topic. Suggestions in no association are irrelevant to their input queries and hence are useless.

### 3. LABELING EXAMPLES

Using our proposed criteria, we hired three assessors to label 17,053 suggestions for 200 queries, which are sampled from a recent one month search log of Bing search engine. We will show some examples to demonstrate the advantages of our new assessment criteria in this section.

<table>
<thead>
<tr>
<th>No.</th>
<th>Suggestion</th>
<th>Utility</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>aol instant messenger</td>
<td>better</td>
<td>same intention</td>
</tr>
<tr>
<td>2</td>
<td>aim messenger</td>
<td>same</td>
<td>same intention</td>
</tr>
<tr>
<td>3</td>
<td>aim instant mess</td>
<td>worse</td>
<td>same intention</td>
</tr>
<tr>
<td>4</td>
<td>msn messenger 2011 beta</td>
<td>worse</td>
<td>peer</td>
</tr>
<tr>
<td>5</td>
<td>yahoo messenger express</td>
<td>worse</td>
<td>peer</td>
</tr>
</tbody>
</table>

Table 1: Suggestions “aol instant mess” and their labels

Table [1] shows the suggestions of the query “aol instant mess” and their labeling results. Although suggestions 1 to 3 are all in the same intention category, 1 can retrieve better documents while 2 and 3 cannot. Therefore, users in scenario 1 would prefer suggestion 1 rather than 2 or 3. Traditional assessments cannot reflect their differences. Suggestion 4 and 5 are peers of the input query. Although their search results do not satisfy the input query’s intention, users in scenario 2 would be interested in them because they help to broaden the knowledge.

<table>
<thead>
<tr>
<th>No.</th>
<th>Suggestion</th>
<th>Utility</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>juicer blender machines</td>
<td>same</td>
<td>peer</td>
</tr>
<tr>
<td>2</td>
<td>juicer machine</td>
<td>same</td>
<td>same intention</td>
</tr>
</tbody>
</table>

Table 2: Suggestions “juicer” and their labels

There are also cases that suggestions not in the same intention category can return better results, such as suggestion 1 in Table [2] and all the suggestions in Table [3]. In these cases, the peer suggestions retrieve better search results because they introduce better query keywords. In traditional assessments, they are likely to be labeled as irrelevant and query suggestions algorithms are discouraged to return them. Therefore, our assessment criteria are more precise at describing the usefulness of query suggestions.

### 4. CONCLUSION

In this paper, we identify two user scenarios of query suggestion and introduce corresponding assessment criteria. The first criterion addresses the scenario that the input query fails to return satisfactory search result. It requires assessors to compare the search results of the suggestions to that of the input query and awards those suggestions having better search results. To the best of our knowledge, this is the first criterion that compares the search result quality of the input query and its suggestions. The second criterion considers different kinds of relationships between an input query and its suggestions. It assumes the input query already returns good search results, and awards the suggestions that help to broaden the user’s knowledge.

Our assessment can be applied to different types of evaluation by assigning a suitable score value to each category depending on the evaluation purpose. Then precision, recall, MAP, and even DCG/NDCG can be applied to measure the effectiveness of a suggestion list.

### 5. REFERENCES


