ANNOTATING SEARCH RESULT USING QUERY BASED ANNOTATOR FROM WEB DATABASE

B.Prema#1, M.Puspalatha#2
#1M.Phil Research scholar, Dept of Computer Science, Padmavani Arts and Science College for Women.
#2Assistant Professor, Dept of Computer Science, Padmavani Arts and Science College for Women

Abstract—Web search applications represent user information needs by submission of query to search engine. But still the entire query submitted to search engine doesn’t satisfy the user information needs, because users may want to get information on diverse aspects when they submit the same query. From this discovering the numeral of dissimilar user search goals for query and depicting each goal with several keywords automatically become complicated. The suggestion and examination of user search goals can be very valuable in improving search engine importance and user knowledge. The numeral of dissimilar user search goals for query by k-means clustering is discovered by user feedback sessions. Pseudo document with k-means clustering is generated by user feedback sessions. Clustering Pseudo documents with k-means clustering results are computationally difficult and semantic similarity between the pseudo terms is also important while clustering. To conquer this problem proposed a FCM (fuzzy c means) clustering algorithm to group the pseudo documents and it also measure the semantic similarity between the pseudo terms in the documents. The FCM algorithm divides pseudo documents data for dissimilar size cluster by using fuzzy systems. FCM choosing cluster size and central point depend on fuzzy model. The FCM clustering algorithm it congregate quickly to a local optimum or grouping of the pseudo documents in well-organized way.

Keywords: User search goals, feedback sessions, pseudo documents, classified precision

I. INTRODUCTION

World Wide Web (WWW) is very popular and interactive. It has become an important source of information and services. The web is huge, diverse and dynamic. Extraction of interesting information from Web data has become more popular and as a result of that web mining has attracted lot of attention in recent time. Web mining is the process of discovering knowledge, such as patterns and relations, from Web data. Web mining generally has been divided into three main areas: content mining, structure mining and usage mining. Each one of these areas are associated mostly, but not exclusively, to these three predominant types of data found in the Web.

Content: The real data that the document was designed to give to its users. In general this data consists mainly of text and multimedia.

Structure: This data describes the organization of the content within the Web. This includes the organization inside a Web page, internal and external links and the website hierarchy.

Usage: This data describes the use of a website or search engine, reflected in the Web server’s access logs, as well as in logs for specific applications.

II. FCM ALGORITHM

1. Initialize U=[u_{ij}(0)] matrix, U^{(0)}
2. At k-step: calculate the centers vectors C^{(k)}=[c_j] with U^{(k)}
   \[ c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \]
3. Update U^{(k1)}, U^{(k+1)}
   \[ u_{ij} = \frac{1}{\sum_{k=1}^{C} \left( \frac{\|x_i - c_k\|^2}{\|x_i - c_j\|^2} \right)^{-\frac{1}{m-1}}} \]
   If \( \|U^{(k+1)} - U^{(k)}\| < \) then STOP; otherwise return to step 2.

III. FUZZY C MEANS CLUSTERING

Fuzzy clustering is a class of algorithms for cluster analysis in which the allocation of data points to clusters is not "hard" (all-or-nothing) but "fuzzy" in the same sense as fuzzy logic. In fuzzy clustering, every point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely too just one cluster. Thus, points on the edge of a cluster may be in the cluster to a lesser degree than points in the center of cluster. An overview and comparison of different fuzzy clustering algorithms is available. Any point x has a set of coefficients giving the degree of being in the kth cluster w_k(x). With fuzzy c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster. The algorithm minimizes intra-cluster variance as well, but has the same problems as k-means; the minimum is a local minimum, and the results depend on the initial choice of weights. Using a mixture of Gaussians along with the expectation-maximization algorithm is a more statistically formalized
method which includes some of these ideas: partial membership in classes. Another algorithm closely related to Fuzzy C-Means is Soft K-means. Fuzzy c-means has been a very important tool for image processing in clustering objects in an image. In the 70's, mathematicians introduced the spatial term into the FCM algorithm to improve the accuracy of clustering under noise.

User Query

Different users may want to get information on different aspects when they submit the same query. For example, when the query “the sun” is submitted to a search engine, some users want to locate the homepage of a United Kingdom newspaper, while some others want to learn the natural knowledge of the sun. Therefore, it is necessary and potential to capture different user search goals in information retrieval. We define user search goals as the information on different aspects of a query that user groups want to obtain. Information need is a user’s particular desire to obtain information to satisfy his/her need. User search goals can be considered as the clusters of information needs for a query. The inference and analysis of user search goals can have a lot of advantages in improving search engine relevance and user experience.

Feedback Sessions

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. We demonstrate that clustering feedback sessions is more efficient than clustering search results or clicked URLs directly. Moreover, the distributions of different user search goals can be obtained conveniently after feedback sessions are clustered. The feedback session contains URL details with view details.

Pseudo Documents

The Pseudo Documents are constructed based on feedback sessions. The Pseudo documents contain all keywords. The keywords related to ambiguous query. We cluster pseudo-documents by Fuzzy C means clustering which is simple and effective. Since we do not know the exact number of user search goals for each query, we set sessions to be different values and perform clustering based on different values, respectively. The optimal value will be determined through the evaluation criterion. The pseudo-documents can enrich the URLs with additional textual contents including the titles and snippets. Based on pseudo-documents, user search goals can then be discovered and depicted with some keywords.

Cap Calculation

In this module implement the novel evaluation criterion classified average precision (CAP) to evaluate the performance of the restructured web search results. CAP is extended version of AP and VAP. In AP and VAP, we can’t analyze the risks. If all the URLs in the search results are categorized into one class, Risk will always be the lowest namely 0; however, VAP could be very low. Generally, categorizing search results into fewer clusters will induce smaller Risk and bigger VAP, and more clusters will result in bigger Risk and smaller VAP. The proposed CAP depends on both of Risk and VAP.

IV. RECONSTRUCTED RESULTS

Search engines always return millions of search results, it is necessary to organize them to make it easier for users to find out what they want. Restructuring web search results is an application of inferring user search goals. We will introduce how to restructure web search results by Inferring user search goals at first. Then, the evaluation based on restructuring web search results. The original search results are restructured based on the user search goals inferred from the user search. Then, we evaluate the performance of restructuring search results by our proposed evaluation criterion CAP. And the evaluation result will be used as the feedback to select the optimal number of user search goals.

V. EXPERIMENTAL RESULTS

The performance is evaluated on the basis of two factors: precision and Recall. The precision and recall is calculated for performance of alignment. The precision for performance of alignment is as follows.

\[
\text{Precision} = \frac{\text{Correctly Aligned Data units}}{\text{Aligned Data Unit}} \times 100
\]

\[
\text{Recall} = \frac{\text{Data units that are Correctly Aligned}}{\text{Manually Aligned Data Unit}} \times 100
\]

The performance of the basic annotator are compared and shown in the table 5.4. The evaluation shows the combination of all annotators give the most accurate result than finding each one individually. Comparing others table annotator gives nearly an accurate result.

<table>
<thead>
<tr>
<th>Domain</th>
<th>BOOK</th>
<th>MOBILE</th>
<th>GAMES</th>
<th>AVERAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>TA</td>
<td>98</td>
<td>96.3</td>
<td>97.7</td>
<td>97</td>
</tr>
<tr>
<td>QA</td>
<td>98.4</td>
<td>97.3</td>
<td>98.7</td>
<td>98</td>
</tr>
<tr>
<td>IA</td>
<td>96.4</td>
<td>95.1</td>
<td>97.7</td>
<td>97.3</td>
</tr>
<tr>
<td>TQI</td>
<td>99.7</td>
<td>99.8</td>
<td>99.2</td>
<td>99.3</td>
</tr>
</tbody>
</table>
The optimal feature weights obtained through the method over data set is \((0.7, 0.9, 1.0)\) and 0.59 for clustering threshold \(T\). The average alignment precision and recall are converged at about 98%. The learning result shows the data type and the presentation style is the most important features in our alignment method. Then, it applies our annotation method on first Dataset to determine the success rate of each annotator.

Table 5.4 shows the performance of our combined annotator method for all 90 pages in Second Dataset. The precision and recall for every domain are above 97%, and the average precision and recall across all domains are above 98%. The performance is consistent with that obtained over the training set. The errors usually happen in the following cases. First, some composite text nodes failed to be split into correct data units when no explicit separators can be recognized.

The experiments data from various domains with respect to three annotators only. The annotators used include table annotator, query-based annotator in-text prefix/suffix annotator. The three annotators are supported by the prototype application and it is extensible so as to support more annotators in future. The performance of data alignment and annotation are presented in table.

As presented in Table 5.4, it is evident that more than 98% precision and recall were recorded for both the performances such as data alignment and annotations. The table also shows the performance of annotation with wrapper. The results are presented in the following graph.

The data will be aligned automatically, it is critical to achieve accurate alignment. The method is used and it is a clustering based shifting method which utilizes automatically accessible features. By this method, it is very much capable of handling different types of relationships between HTML text nodes and data units.

The experiments data from various domains with respect to three annotators only. The annotators used include table annotator, query-based annotator in-text prefix/suffix

Figure: 1 it is evident that the prototype application is capable of producing annotations automatically given search results. The performance of the application is encouraging and the application can be used in the real world applications. It is evident that more than 98% precision and recall were recorded for both the performances such as data alignment and annotations.

VI. CONCLUSION

This work, group the content into different category related to what we are searching about and also provides data unit level annotation which means order or group the content which belongs to our wish. A user-based evaluation is performed, in order to demonstrate the effectiveness of the automatic annotation method. Moreover, a comparative evaluation validates that, the proposed hybrid search, outperforms in all cases the keyword-based and semantic based search in terms of precision and recall. Finally, all the proposed methods are implemented as a fully functional
tool. Our experimental results show that the precision and recall of this method are both above 98 percent.

REFERENCES


