International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Search Result Optimization using Annotators

Vishal A. Kamble[1], Amit B. Chougule[2]

[1] Department of Computer Science and Engineering, D Y Patil College of engineering, Kolhapur, Maharashtra, India

[2]Professor, Department of Computer Science and Engineering, D Y Patil College of engineering, Kolhapur, Maharashtra, India

Abstract: ***An increasing number of databases have become web accessible through HTML form-based search interfaces. The dataunits returned from the underlying database are usually encoded into the result pages dynamically for human browsing. For the encoded data units to be machine processable, which is essential for many applications such as deep web data collection and Internet comparison shopping, they need to be extracted out and assigned meaningful labels. In this paper, we present an automatic annotation approach that first aligns the data units on a result page into different groups such that the data in the same group have the same semantic. Then, for each group we annotate it from different aspects and aggregate the different annotations to predict a final annotation label for it. An annotation wrapper for the search site is automatically constructed and can be used to annotate new result pages from the same web database. Our experiments indicate that the proposed approach is highly effective.***

Keywords: Web Mining, Clustering, Optimization, Annotators, Search engine.

1. Introduction

Now days, people are accessing a lot of information through the World Wide Web. For accessing the information they are using search engines like Google, Yahoo etc. Search engine produces the results obtained through query given by people. This result is effect of query executed on Web Databases. This provides the information needed or requested by the people along with partially related information. The produced information contains number of Search Result Records (SRRs). Where, each SRR contains multiple data units which describe one aspect of a real-world entity. Each data unit corresponds to the value of a record under an attribute.

It becomes crucial to identify relevant and irrelevant information from SRRs to get the required information. Annotation is used to identify relevant information. An annotation is metadata (e.g. a comment, explanation, and presentational markup) attached to text, image, or other data.

2. Literature Review

2.1 Annotating Search Results from Web Databases

Authors have represented an automatic annotation approach which first aligns the data units on a result page into different groups such that the data in the same group have the same semantics. Then, for each group the author annotates it from different aspects and aggregates the different annotations to predict a final annotation label for it [1].

2.2 Review on automatic Annotation search from Web databases

Authors presented an automatic annotation approach. The motivation behind such systems lies in the emerging need for going beyond the concept of "human browsing". The World Wide Web is today the main "all kind of information" repository and has been so far very successful in disseminating information to humans [2]. .

2.3 A survey on annotating search results from web databases

Authors introduces the automatically annotation wrapper generation mechanism used for annotate new result records from the same web database [3].

2.4 Application Research of k-means Clustering Algorithm in Image Retrieval System

Authors suggest that, Image retrieval algorithms always use the similarity between the query image and images in image database. However, they ignore the similarities between images in image database. Then, authors addressed this problem by introducing a graph-theoretic approach for image retrieval post-processing step by finding image similarity clustering to reduce the images retrieving space [4].

2.5 Random Indexing Approach for Web User Clustering and Web Pre-fetching

Author focuses on discovering latent factors of user browsing behaviors based on Random Indexing and detecting clusters of Web users according to their activity patterns acquired from access logs. Experiments are conducted to investigate the performance of Random Indexing in Web user clustering tasks. The experimental results show that the proposed RI-based Web user clustering approach could be used to detect more compact and well-separated user groups than previous approaches [5].

2.6 Introduction to Search Engine Optimization

Web pages that contain the words that your target audience is typing into search queries generally have greater search engine visibility than pages that contain little or no keywords. The way your web pages are linked to each other also affects your site's search engine visibility. If search engine spiders can find your pages quickly and easily, your site has a much better chance of appearing at the top of search results. If two web sites have the same text component and link component

"weights," the site that end users click the most will usually rank higher. Sometimes, a popular web site will consistently rank higher than sites that use plenty of keywords. Therefore, building a site that appeal to both directory editors and your target audience is very important for maximum search engine visibility [6].

2.7 Annotating Search Results from Web Databases Using Clustering-Based Shifting

Authors have represented an automatic annotation approach that first aligns the data units on a result page into different groups such that the data in the same group have the same semantics. Then, for each group authors annotate it from different aspects and aggregate the different annotations to predict a final annotation label. An annotation wrapper for the search site is automatically constructed and can be used to annotate new result pages from the same web database [7].

3. Problem Definition

There is high demand for collecting data of interest from multiple WDB's [1]. As a result, it produces number of Search Result Records (SRRs). The produced SRRs will contain some irrelevant information which user is not interested in. The users are interested in only specific and relevant information.

Our objective is to provide a mechanism which will attempt to optimize the search results by removing irrelative information and provide most relevant information to the users.

Our proposed system uses the automatic annotation technique [1], automatic annotation wrapper [3] and clustering k-means algorithm for grouping [4], random indexing technique [5] for ranking purpose, search engine optimization technique [6], and the concept of probabilistic labeling [7] to achieve our objective.

4. System Design and Implementation

4.1 Scope

Search Result Optimization using Annotators (SROA) will offer a mechanism to identify the specific information by applying annotation and optimization. We can achieve the following:

1. Users will search any content in a search engine.
2. It will group the result obtained by search engine into different category and provide data unit level annotation.
3. Then, it will optimize the retrieved search result records (SRR) by making use of Random Indexing (RI) and Similarity Measure.

SROA can be applied in the applications like getting specific news from web pages.

4.2 System Architecture

Figure 1 shows the system architecture, which shows the mechanism to optimize the search by removing irrelative information and provide most relevant information to the users. This works as follows:

- First, user will search needed information by providing related keywords and respective category.
- Then it will extract the relevant information from the results obtained. This will generate a simplified HTML or XML source corresponding to SRR.
- Then, it will align the data units and organize it with different groups having same semantic. K-means algorithm will be used for clustering purpose.
- After that annotators will be used for annotating the SRRs.
- Then, the SRRs will be optimized by using Similarity Measures and Random Indexing (RI) calculation which will produce optimized results.

Our proposed system is being categorized into following modules:

- Information Extraction (IE) Module
- Pre Processing and Alignment Module
- Clustering of aligned data Module
- Annotation and Optimization of Clustered data Module
- Analysis of the system

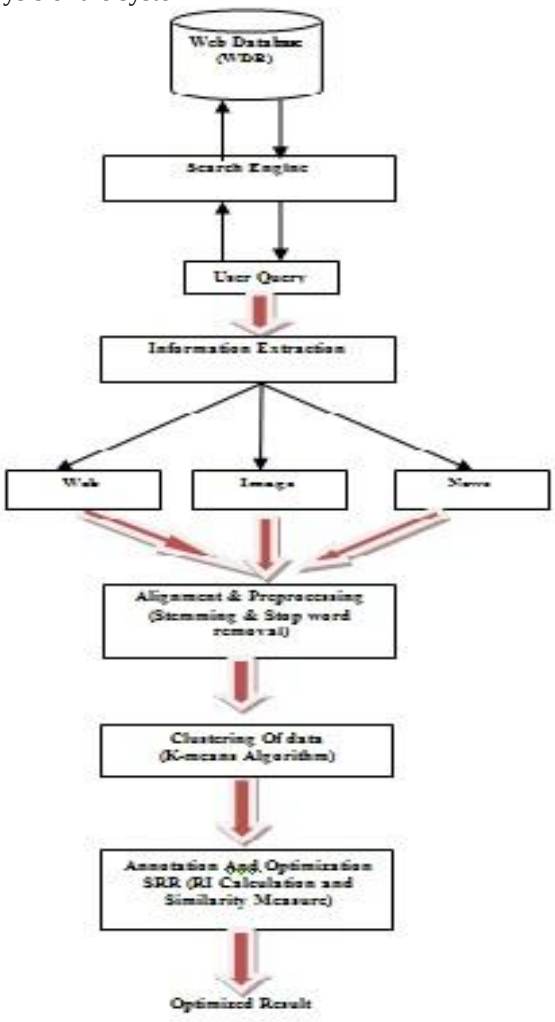

Figure 1: System Architecture

4.2.1 Information Extraction (IE) Module

In this module we will extract contents from result page returned from WDB. The result page contains multiple search result records (SRR). Each SRR have number of data units. Each data unit describes semantic of an entity.
Then we will form relational database with each rows representing search result records.

4.2.2 Pre Processing and Alignment Module

In this module, we will identify all data units in the SRRs and then organize them into different groups. Each group will correspond to the different concept. Then, we will apply alignment algorithm to align data units of same concepts into each column across all SRRs. The alignment algorithm will be as follows:

```
ALIGN (SRRs)
   J←1;
   While true
//Create alignment groups
   For i←1 to number of SRRs
      Gj←SRR[i][j]
   If Gj is empty
      Exit;
   V←CLUSTERING(G);
   If |V| > 1
//Collect all data units in groups following j
   S←Ø
   For x←1 to number of SRRs
        For y←j+1 to SRR[i].length
      S←SRR[x][y];
//Find cluster c least similar to following groups
   V[C]=min(sim(V[k],S));    min from k to |V|
//Shifting
   For k←1 to |V| and k ≠ c
       For each SRR[x][j] in V[k]
      Insert NIL at position j in SRR[x];
   j←j+1     //Move to next group
```

4.2.3 Clustering of aligned data Module

In this module, we will use a clustering-based shifting technique to align data units into different groups so that the data units inside the same group have the same semantic. The grouping data units of the same semantic will help to identify the common patterns and features among these data units.
For this, we will apply K-means clustering algorithm which works as follows:

1) Select an initial partition with K=k clusters; repeat steps 2 and 3 until cluster membership stabilizes.
2) Generate a new partition by assigning each pattern to its closest cluster center.
3) Compute new cluster centers.
4) Instead of considering only the DOM tree or other HTML tag tree structures of the SRRs, we will also consider other important features like presentation style, data type etc. shared among data units to align the data units. So that data in one cluster will have most similar data.

4.2.4 Annotation and Optimization of Clustered data Module

In this module, we will apply query based annotator, Schema value annotator and frequency based annotator to produce a label for units within their specific group. Also, we will use a probability model to determine the most appropriate label for each group. Then, we will generate annotation rule which describes how to extract the data units in the result page and what the appropriate semantic label should be. The rules for all aligned groups, collectively, form the annotation wrapper for the corresponding WDB. Annotation wrapper can be used to directly annotate the data retrieved from the same WDB in response to new queries without the need to perform the alignment and annotation phase again. Annotation wrappers are essential for online applications as it can perform annotation quickly.

After annotating we will optimize search results records by making use of Random indexing and similarity measure techniques which are described as follows:

Random Indexing (RI)

The basic idea of Random Indexing is to accumulate context vectors based on the occurrence of words in contexts. This technique can be used with any type of linguistic context, is inherently incremental, and does not require a separate dimension reduction phase. The Random Indexing technique can be described as a two-step operation.

Step1: A unique d-dimensional index vector is assigned and randomly generated to each context.

Step2: Context vectors are produced by scanning through the text.

Similarity Measures

Similarity/distance measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. In many cases, these characteristics are dependent on the data or the problem context at hand, and there is no measure that is universally best for all kinds of clustering problems.
Similarity/distance measures map the distance or similarity between the symbolic descriptions of two objects into a single numeric value, which depends on two factors:

1. The properties of the two objects and
2. The measure itself

4.2.5 Analysis of the system

It will attempt to compare the results produced by traditional search engine with results produced by 'search results optimization using annotators'. For this we will consider precision and recall as parameters.

5. Experimental Result and Analysis

Work is carried out using C#.net on Visual Studio 10(4.0) and back end used is SQL server 2005.

The following Table 1 shows result comparison of Google Search Engine and SROA System. As compared our system with Google Search Engine we observed that for given user query for example "computer" user keyword Google search

Engine gives 62 links and our SROA system gives 13 links. These 13 links are optimized and relevant to user query.

Table 1: Comparisons of Google Search Engine and SROA System

User Query	*Google Search Engine (Search Result in no. of links)*	*SROA System (Search Result in no. of links)*
Computer	62	13
Arun jetali	59	19

Figure 2 shows graphical representation of comparison between Google Search Engine and SROA System

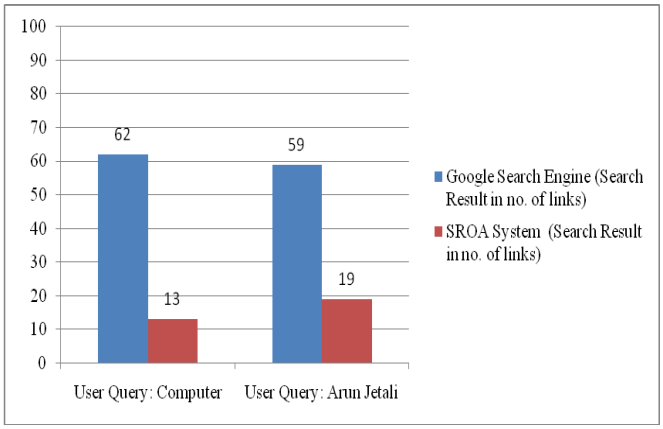

Figure 2: Graphical Representations of comparisons between Google Search Engine and SROA System

The following Table 2 shows result comparison of Google Search Engine for news category and SROA System.

Table 2: Comparisons of Google Search Engine for news and SROA System

User Query	Google Search Engine for News (Search Result in no. of links)	SROA System for News (Search Result in no. of links)
Salman Khan	72	8
Jailalita	43	4

Figure 3 shows graphical representation of comparison between Google Search Engine for news category and SROA System.

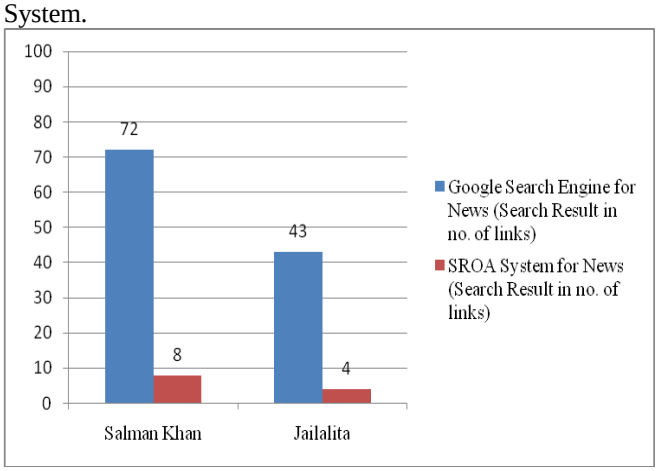

Figure 3: Graphical Representations of comparisons between Google Search Engine for news and SROA System.

The following Table 3 shows result comparison of Yahoo Search Engine and SROA System. As compared our system with Yahoo Search Engine we observed that for given user query for example "computer" user keyword Google search Engine gives 39 links and our SROA system gives 10 links. These 10 links are optimized and relevant to user query.

T

able 3: Comparisons of Yahoo Search Engine and SROA System

User Query	*Yahoo Search Engine (Search Result in no. of links)*	*SROA System (Search Result in no. of links)*
Computer	39	10
Arun jetali	42	7

Figure 4 shows graphical representation of comparison between Yahoo Search Engine and SROA System.

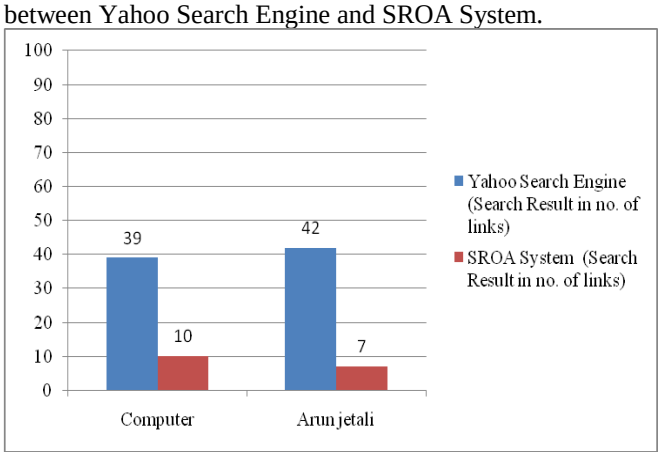

Figure 4: Graphical Representations of comparisons between Yahoo Search Engine and SROA System

The following Table 4 shows result comparison of Bing Search Engine and SROA System. As compared our system with Bing Search Engine we observed that for given user query for example "Samsung" user keyword Google search Engine gives 58 links and our SROA system gives 9 links. These 9 links are optimized and relevant to user query.

Table 4: Comparisons of Bing Search Engine and SROA System

User Query	Bing Search Engine (Search Result in no. of links)	SROA System (Search Result in no. of links)
Jailalita	67	12
Samsung	58	9
Shivaji Maharaj	46	5

Figure 5 shows graphical representation of comparison between Bing Search Engine and SROA System.

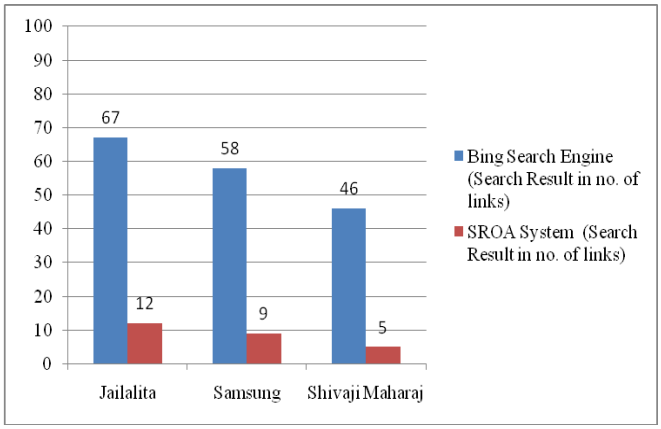

Figure 5: Graphical Representations of comparisons between Bing Search Engine and SROA System

6. Conclusion

In this paper, we are searching the specific results from large web databases using the three annotators. In this work first we parse the web pages and store to the databases. We first identify all the data units available and organize them in to different groups. Grouping data units of same semantics helps to identify the common patterns between these data units which are the basis for annotator.

In next phase many basic annotators are introduced in which some common features are followed. Each basic annotator is used to label the units of same group. It is also used for identifying most appropriate label for each group.

The next phase is the annotation wrapper generation phase. In this phase an annotation rule is generated for each identified concepts which shows how to extract data units of same group. The collective annotation rule for an aligned group is known as annotation wrapper for the corresponding web database. We apply optimization techniques on results obtained from annotation which removes unwanted links and gives relevant link to the user query.

6.1 Future Work

The scope of the work is, when we search any content in a search engine, it will 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 Searched Query.

In online shopping sites databases are in unorganized manner, it is quite difficult to users. To overcome this problem automatic multi-annotator approach is proposed. There is still room for improvement in the automatic annotator which makes the annotator dynamic. In the dynamic annotator we can add features by using powerful classification technique (more advanced classification algorithms).

This work can be carried forward for optimization of Images, Video type of contents SRR with applying slightly different filtering techniques.

References

[1] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Men "Annotating Search Results from Web Databases" IEEE 2013.
[2] Kiran C. Kulkarni , S. M. Rokade "Review On Automatic Annotation from search from Web Database " International Journal of Emerging Technology and Advanced Engineering 2014
[3] Y. Pauline Jeba, Mrs. P. Rebecca Sandra " Survey on Annotating search results from Web", International Journal Of research in computer application and Robotics Database 2013
[4] Hong Liu 1, and Xiaohong Yu "Application Research of k-means Clustering Algorithm in Image Retrieval System" Proceedings of the Second Symposium International Computer Science and Computational Technology 2009
[5] Miao Wan, Arne J¨onsson, Cong Wang, Lixiang Li, and Yixian Yang "A Random Indexing Approach for Web User Clustering and Web Prefetching".
[6] Introduction to Search Engine Optimization
[7] Saranya.J1, SelvaKumar.M2, Vigneshwaran.S3, Danessh. M.S "Annotating Search Results from Web Databases Using Clustering-Based Shifting" International Journal of Innovative Research in Science, Engineering and Technology,2014

Author Profile

Mr. V. A. Kamble , BE degree in Computer Science and engineering from Shivaji University Kolhpur, Maharashtra, India.Currently he is pursuing his ME in Computer Science and Engineering in D.Y.Patil College of Engineering and Technolgy,Kolhapur,Maharashtra and working as a Assistant Professor at Annasaheb Dange college of engineering,Ashta,Tal:Walwa,Dist Sangli.

Prof. A. B. Chougule, M. Tech. in Computer Science and Technology from Shivaji University Kolhapur, Maharashtra, India. Working as Professor and Head of Computer Science and engineering Department at Bharati Vidyapeet's College of Engineering, Kolhapur, Maharashtra, India.