Notice of Violation of IEEE Publication Principles

"Language Models for Web Object Retrieval,"
by Jianfeng Zheng; Zaiqing Nie
in the Proceedings of the International Conference on New Trends in Information and Service Science, (NISS), June 2009, pp. 282-287

After careful and considered review of the content and authorship of this paper by a duly constituted expert committee, this paper has been found to be in violation of IEEE's Publication Principles.

This paper is a verbatim copy of the paper cited below. The lead author, Jianfeng Zheng, submitted the copied paper without the knowledge or permission of the coauthor, Zaiqing Nie.

Due to the nature of this violation, reasonable effort should be made to remove all past references to this paper, and future references should be made to the following article:

"Web Object Retrieval"
by Zaiqing Nie, Yunxiao Ma, Shuming Shi, Ji-Rong Wen, and Wei-Ying Ma
in the Proceedings of the 16th International World Wide Web Conference (WWW2007), May 2007, ACM

2009 International Conference on New Trends in Information and Service Science

Language Models for Web Object Retrieval

Jianfeng Zheng
School of Economics and Management
BUPT
Beijing,China
kezheng@microsoft.com

Zaiqing Nie
Microsoft Research Asia
Microsoft
Beijing, China
znie@microsoft.com

Abstract

Document-level information retrieval can unfortunately lead to highly inaccurate relevance ranking in answering object-oriented queries. A paradigm is proposed to enable searching at the object level. However, this reliability assumption is no longer valid in the object retrieval context when multiple copies of information about the same object typically exist. To resolve multiple copies inconsistent issue, we propose several language models for Web object retrieval, namely an unstructured object retrieval model, a structured object retrieval model, and a hybrid model with both structured and unstructured retrieval features. We test these models on a paper search engine and compare their performances. We conclude that the hybrid model is the superior by taking into account the extraction errors at varying levels.

Keywords: Web Objects, Information Retrieval, Language Model, Information Extraction

1. Web Object and Object Extraction

Figure 1 shows the compounds of a Web object and a flowchart to extract the object from Web sources. The key messages conveyed by the figure are:

- The contents of a Web object are aggregated from multiple Web sources. These copies may be inconsistent because of the diverse Web site qualities and the limited performance of current information extraction techniques.

- From each source, two steps are taken to extract the wanted information. First, record extraction [6] is applied to get data records relevant to the domain from the resource. Second, attribute extraction [12] is used to label different portions of each extracted record as different attributes. Both of the two steps are unlikely to be accurate. Record extraction can extract a totally wrong record, miss some parts of a record, or add irrelevant information to a record. Attribute extraction may wrongly label an attribute or not identify an attribute. But, in practice, the accuracy of every extraction algorithm on each Web source can be reasonably measured by using some test dataset. Therefore, we can assign the accuracy number to each extraction function in the figure and take it as a quality measurement of the data extracted. We use a_k to denote the accuracy of record detection, and γ_k to denote the accuracy of attribute extraction of record *k.*

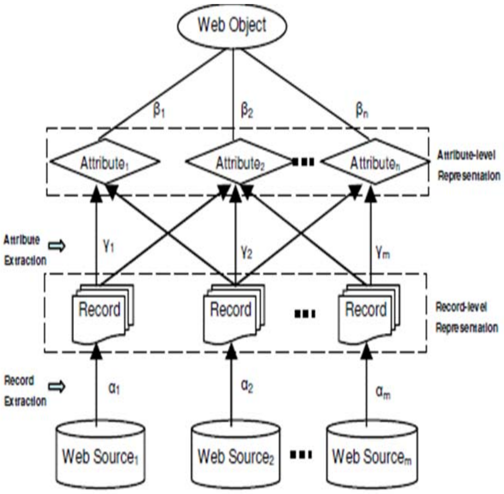

Figure 1. Web Object and Object Extraction

- An object can be described at two different levels. The first one is the record-level representations, in which an object can be viewed as the collection of a set of extracted records and the attributes of each record are not further distinguished. The second on is the attribute-level representations, in which an object is made up of a set of attributes and each attribute is a collection of attribute

978-0-7695-3687-3/09 $25.00 © 2009 IEEE
DOI 10.1109/NISS.2009.21

instances extracted from the records in multiple sources.

- The importance of the j^{th} attribute β_j , indicates the importance level of the attribute in calculating relevance probability. The problem of using differing weights for different attributes has been well studied in existing structured document retrieval work [12][8] and can be directly used in our Web object retrieval scenario.

2. Web Object Retrieval

Our goal in this paper is to explore effective models to retrieval Web objects described above. The retrieval models should be insensitive to data errors and can achieve stable performance for data with varying extraction accuracy. In document-level information retrieval, there is no concept of correctness. This is because there is no pre-defined semantic meaning of a document, and all the words and sentences in the document will define the meaning of the document. However the meaning of real world objects is pre-defined and the descriptions about the objects on the Web may be incorrect. Since the users usually want to see the correct information about the most relevant real-world objects first, it is critical to be able to use the accuracy of the extracted object descriptions in calculating the relevance probabilities of their corresponding real-world objects.

3. Language Models

In this section, we present a language model to estimate the relevance between an object and a query. We first provide background on language modeling for document retrieval. We then propose several language models for Web object retrieval, namely an unstructured object retrieval model, a structured object retrieval model, and a hybrid model with both structured and unstructured retrieval features.

3.1. Background on Language Modeling

Language models interpret the relevance between a document and a query as the probability of generating the query from the document's model. That is,

$$P(D \mid Q) \propto P(Q \mid D) \cdot P(D)$$

For a query Q, if independence among query terms are assumed, then it can be proved (by simple probability calculations) that,

$$P(Q \mid D) = \prod_{i=1}^{|Q|} P(w_i \mid D)$$

Where w_i is the i^{th} query term of Q, |Q| is denoted as the length of Q, and is the $P(w_i \mid D)$ probability of generating term wi from the language model of D.

Given word w and document D, maximum likelihood estimation (MLE) is commonly used to estimate probability P(w|D). Smoothing, which adjusts term probabilities to overcome data sparseness, is critical to the performance of language models. Among various smoothing methods, the Dirichlet prior smoothing is frequently discussed. By maximum likelihood estimation and Dirichlet smoothing, the probability of generating term w by the language model of document D can be estimated as follows,

$$P(w \mid D) = \lambda \frac{tf(w,D)}{|D|} + (1-\lambda)\frac{tf(w,C)}{|C|}$$

where |D| is the length of document D, tf(w,D) is the term frequency (i.e. number of terms) of term w in D, |C| is the number of terms in the whole collection, and tf(w,C) is the term frequency of term w in the whole collection C. In the above formula, λ can be treated as a parameter with its value in [0, 1]. It is common to let λ rely on document length |D|, as follows,

$$\lambda = \frac{|D|}{|D| + \mu}$$

where μ is a parameter and it is common to set it according to the average document length in the collection.

3.2. Web Object Retrieval

In the following subsections, we present language models for Web object retrieval.

3.2.1. Bag of Words (BW)

In this model, we treat all term occurrences in a record equally and there is no difference between records either. This is actually the traditional document retrieval model that considers all the information about the same object as a bag of words. Indeed, this is a special case for the record-level.

3.2.2. Unstructured Object Retrieval (UOR)

One simple way of scoring a Web object against a query is to consider each record as the minimum retrieval unit. In this way, all the information within a record is considered as a bag of words without further differentiating the attribute values of the object, and we only need to know the accuracy of record extraction. The advantage of this model is that no attribute value extraction is needed, so we can avoid amplifying the attribute extraction error for some irregular records whose information cannot be accurately extracted. This model can also be called unstructured object retrieval model since it treats each record as an unstructured document.

Now we present a language model for record-level Web object retrieval. If we consider all the information about an object as a big document consisting of K records, we can have a language model for each record and combine them, as [8] have been done. One approach to combining the language models for all the records of object o is as follows,

$$P(w \mid o) = \sum_{k=1}^{k} (a_k P(w \mid R_k))$$

where is the $P(w \mid R_k)$ probability of generating w by the record R_k ,and . is a_k the accuracy of record extraction. $P(w \mid R_k)$ can be computed by treat each record R_k as a document,

$$P(w \mid R_k) = \lambda \frac{tf(w, R_k)}{\mid R_k \mid} + (1-\lambda) \frac{tf(w, C)}{\mid C \mid}$$

Where C is the collection of all the records, and is set according to Dirichlet prior smoothing.

In this model, we only need to know the record extraction accuracy which can be easily obtained through empirical evaluation. Note that the parameters a_k are normalized accuracy numbers and 1 k

$$\sum_{k} a_k = 1$$

The intuition behind this model is that we consider all the fields within a record equally important and give more weight to the correctly detected records.

3.2.3. Multiple Weighted Fields (MWF)

This method assigns a weight to each attribute (β_j) and amends the $P(w \mid o_{jk})$ by multiplying the weight of the corresponding attribute. However, it does consider the extraction error. We use the same a_k and γ_k . for all records in the attribute-level representation model for this model

3.2.4. Structured Object Retrieval (SOR)

For the object records with good extraction patterns, we do hope to use the structural information of the object to estimate relevance. It has been shown that if we can correctly segment a document into multiple weighted fields (i.e. attributes), we can achieve more desirable precision [12][8]. In order to consider the weight difference of different fields and avoid amplifying the attribute extraction error too much, we need to consider attribute extraction accuracy. This model can also be called structured object retrieval model since it treats each record as a structured document. We consider all the information about an object as a big document consisting of K records and each record has M fields (i.e. attributes), and we use the formula below to estimate the probability of generating term w by the language model of object,

$$P(w \mid o) = \sum_{K=1}^{K} (a_k \gamma_k \sum_{j=1}^{M} \beta_k P(w \mid o_{jk}))$$

Where $a_k \gamma_k$ together can be considered as the normalized accuracy of both record detection and attribute extraction of record k , and $a_k \gamma_k = 1$ is the importance of the j^{th} field, and $\sum_{j} \beta_j = 1$ .Here $P(w \mid o_{jk})$ is the probability of generating w by the j^{th} field of record k. $P(w \mid o_{jk})$ can be computed by treating each o_{jk} as a document,

$$P(w \mid o_{jk}) = \lambda \frac{tf(w, o_{jk})}{\mid o_{jk} \mid} + (1-\lambda) \frac{tf(w, C_j)}{\mid C_j \mid}$$

Where C_j is the collection of all the j^{th} fields of all the objects in the object warehouse, and is set according to Dirichlet prior smoothing.

3.2.5. Structured and Unstructured Retrieval (BSUR)

As we discussed earlier, the unstructured object retrieval method has the advantage of handling records with irregular patterns at the expenses of ignoring the structure information, while attributelevel retrieval method can take the advantage of structure information at the risk of amplifying extraction error. We argue that the best way of scoring Web objects is to use the accuracy of extracted object information as the parameter to find the balance between structured and unstructured ways of scoring the objects. We use the formula below to estimate the probability of generating term w by the language model of object,

$$P(w|o)=\sum_{K=1}^{K}(a_k\sum_{j=1}^{M}(\gamma_k\beta_j+(1-\gamma_k)\frac{1}{M}P(w|o_{jk}))$$

The basic intuition behind this formula is that we give different weights to individual fields for correctly extracted records and give the same weight to all the fields for the incorrectly extracted records.

4. Parameter Setting

Below we will use Libra (http://libra.msra.cn), a working scientific Web search engine we have built to motivate the need for object level Web search and its advantages and challenges over existing search engines.

Compared to the traditional unstructured document retrieval, in our model we set a weight of each attribute (β_j). The weights of the attributes are tuned manually by considering the importance of attributes. To determine the extraction accuracy a_k and γ_k , we sampled some data for each data source, then compute the accuracy for both record and attribute extraction results. Table 3 shows the results.

Table 1. Extraction Accuracy Parameters

	Citeseer	ACM	DBLP	SCI	PEv1	PEv2	PEv3
a_k	0.80	0.92	0.96	0.94	0.68	0.69	0.76
γ_k	0.74	0.95	0.97	0.91	0.63	0.73	0.78

Although ACM, DBLP and SCI are built manually and got high extraction accuracy, we can't totally depend on them to ensure data coverage. For example, the ACM only provides about 300,000 papers and many important articles are not covered. In addition, to keep the up to date data, the search engine has to crawl PDFs from the Web and extract info in them. Therefore, we have to utilize information from every source. Because each source provides only a subset of the papers in Libra, no single data source can dominate the results.

5. Experiment Results

For each query, we try the five models over all the information from 7 data sources (DBLP, ACM Digital Library, CiteSeer, SCI, PEv1, PEv2 and PEv3). Then the top 30 results of every query are collected from each algorithm and labeled with relevance judgments. In order to ensure a fair labeling process, all the top papers from all the models are merged before they were sent to the labeler. In this way the labeler could not know the ranked position and the connection between the models and the ranking results. We ask labelers with different background to handle the queries they are familiar with. We observe the precision at 10, precision at 30, average precision (MAP) and the precision-recall curve to measure the performance of all five models. The result clearly shows that the Balancing Structured and Unstructured Retrieval (BSUR) model is consistently better than other models.

In Figure 2 we show the precision at rank=10 of the results returned by the five retrieval models, in Figure 3 we show the precision at rank=30 of the results returned by the five retrieval models, and Figure 4 is the average precision (MAP) for all the five models. The Precision-Recall curve is also plotted in Figure 5. As we can see, the models that considered accuracy levels of the extractors have better precision, and the BSUR model is much better than the other models. This is especially true if we want to reduce the error for the top ranked results (for example, at rank=10).

In addition to the performance test, statistical tests are also used to determine the significance of differences [11][14]. We did the paired t-test analysis on F1 score. After grouped models with insignificant performance, the p-value shows that BSUR is significantly better than {UOR, MWF, SOR} which are significantly better than BW.

We believe that even though several low quality data sources were used, we can achieve good retrieval results by combining all evidence from all data sources. To verify this, each time we use one of our developed extractors (PEv1, PEv2, and PEv3)), and the four Web databases (ACM, Citeseer, DBLP, SCI) to complete our experiments, the quality of PEv1, PEv2 and PEv3 become better and better. The MAP results for the five models are shown in Figure 6.

Because the results of P@10 and P@30 are similar to the MAP results, we omitted them. The result clearly illustrates that the BSUR model is almost insensitive to noise from low quality data sources if we use the evidence from other data sources, and our BSUR model is rather robust. In addition, models that consider extraction accuracy levels are consistently better than comparative models. Finally, the gap between models that consider extraction accuracy and models that not consider extraction accuracy will increase when noise increases.

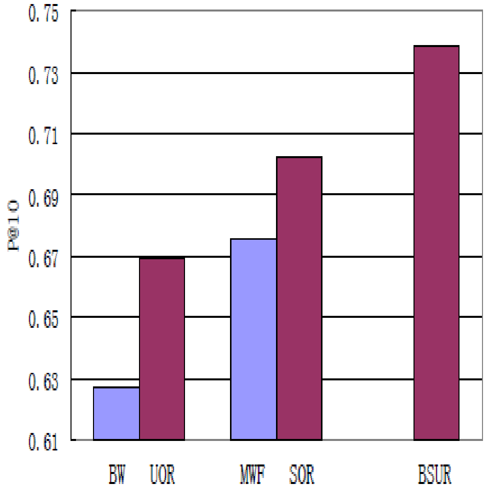

Figure 2. Precision at 10

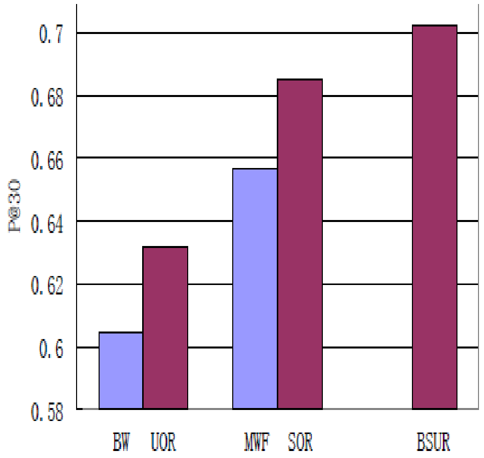

Figure 3. Precision at 30

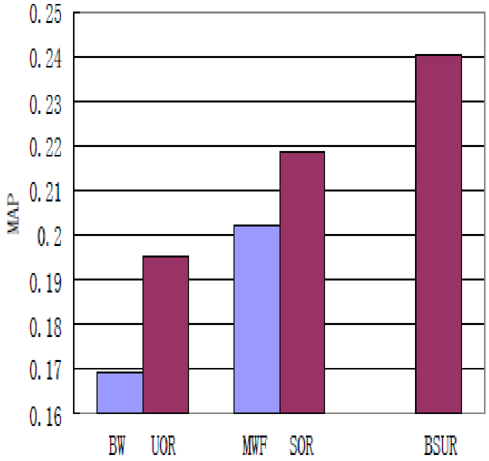

Figure 4. Average Precision (MAP)

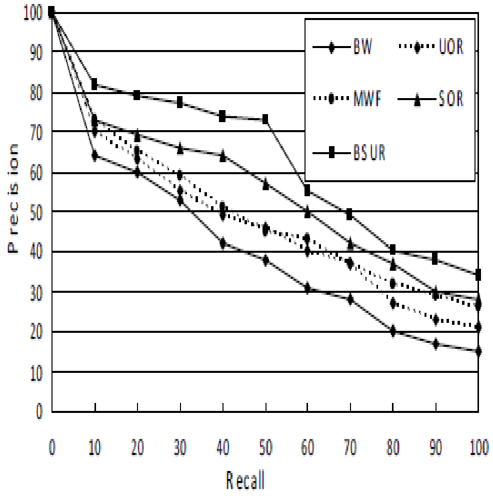

Figure 5. Precision at 11 Standard Recall Levels

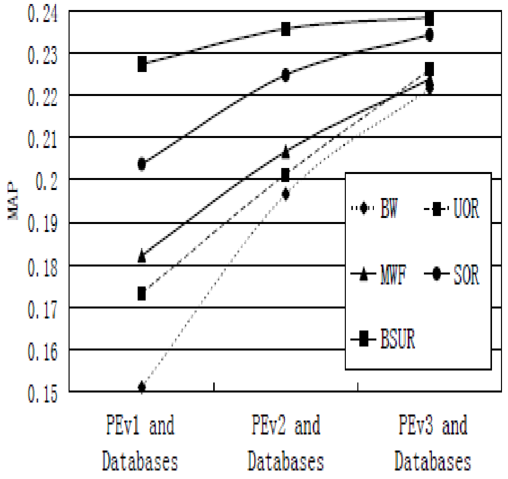

Figure 6. Average Precision (MAP) with Different Quality Data Sources

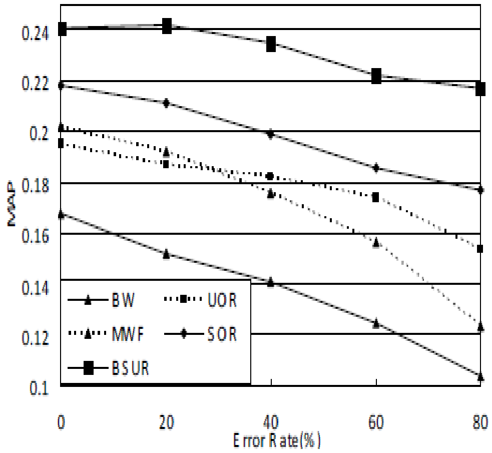

Figure 7. Average Precision (MAP) with Different Error Rate

To better control the error rate of data, we also manually add noise into the dataset. Both of record and attribute level errors of a record are brought in by adding irrelevant words, discarding some words or exchanging words between attributes according to some desired error rate. In this experiment, we introduce noise into ACM and SCI dataset, because they provide full documents data with best quality. The accuracy of these sources are set based on the error rate. Figure 7 shows the MAP results of all the models with different error rates. Because there is much more noise, the improvement and robustness of the model considering data qualities are much more significant.

6. Conclusion

There is lots of structured information about real-world objects embedded in static Web pages or online Web databases. Our work focuses on object level retrieval, which is a completely new perspective, and differs significantly from the existing structured document retrieval and passage/block retrieval work. We propose several language models for Web object retrieval, namely an unstructured object retrieval model, a structured object retrieval model, and a hybrid model with both structured and unstructured retrieval features. We test these models on Libra Academic Search and compare their performances. We conclude that the hybrid model is the superior by taking into account the extraction errors at varying levels.

7. References

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley Publishers, 1999.

[2] Deng Cai, Xiaofei He, Ji-Rong Wen, and Wei-Ying Ma. Block-Level Link Analysis. In Proceedings of SIGIR, 2004.

[3] J. P. Callan. Passage-Level Evidence in Document Retrieval. In Proceedings of SIGIR, 1994.

[4] J.P. Callan. Distributed information retrieval. In Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval, edited by W. Bruce Croft. Kluwer Academic Publisher, pp. 127-150, 2000.

[5] Abdur Chowdhury, Mohammed Aljlayl, Eric Jensen, Steve Beitzel, David Grossman and Ophir Frieder. Linear Combinations Based on Document Structure and Varied Stemming for Arabic Retrieval. In The Eleventh Text REtrieval Conference (TREC 2002), 2003.

[6] Charles L.A. Clarke. Controlling Overlap in Content-Oriented XML Retrieval. In Proceedings of the SIGIR,2005.

[7] Nick Craswell, David Hawking and Trystan Upstill. TREC12 Web and Interactive Tracks at CSIRO. In The Twelfth Text Retrieval Conference(TREC 2003), 2004.

[8] Ronald Fagin, Ravi Kumar, Kevin S. McCurley, Jasmine Novak, D. Sivakumar, John A. Tomlin and David P.Williamson. Searching the Workplace Web. In Proceedings of the Twelfth International World Wide Web Conference, 2003.

[9] Hui Fang, Tao Tao and ChengXiang Zhai. A Formal Study of Information Retrieval Heuristics. In Proceedings of SIGIR, 2004.

[10] David Hull. Using Statistical Testing in the Evaluation of Retrieval Experiments. In Proceedings of the ACM SIGIR, 1993.

[11] Zaiqing Nie, Yuanzhi Zhang, Ji-Rong Wen and Wei-Ying Ma. Object-Level Ranking: Bringing Order to Web Objects. In Proceedings of the 14th international World Wide Web Conference (WWW), 2005.

[12] Zaiqing Nie, Ji-Rong Wen and Wei-Ying Ma. Object-Level Vertical Search. To appear by the Third Biennial Conference on Innovative Data Systems Research (CIDR),2007.

[13] Norbert Fuhr. Probabilistic Models in Information Retrieval. The computer Journal, Vol.35, No.3, pp. 243-255.

[14] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In Proceedings of the ACM SIGIR, 1999.

[15] Jun Zhu, Zaiqing Nie, Ji-Rong Wen, Bo Zhang, Wei-Ying Ma. 2D Conditional Random Fields for Web Information Extraction. In Proceedings of the 22nd International Conference on Machine Learning (ICML), 2005.

[16] Jun Zhu, Zaiqing Nie, Ji-Rong Wen, Bo Zhang, Wei-Ying Ma. Simultaneous Record Detection and Attribute Labeling in Web Data Extraction. ACM Discovery and Data Mining (KDD), 2007.

[17] Zuobing Xu, Ram Akella, Active Relevance Feedback for Difficult Queries. To be published in Proceedings of ACM 17th Conference on Information and Knowledge Management (CIKM) 2008

[18] Zuobing Xu, Ram Akella, Bayesian Logistic Regression Model for Active Relevance Feedback. In Proceedings of the 31st ACM SIGIR Conference, 2008.

[19] Zuobing Xu, Ram Akella, New Probabilistic Retrieval Model Based on the Dirichlet Compound Multinomial Distribution In Proceedings of the 31st SIGIR Conference, 2008.