Survey On Web Object Retrieval Techniques

Sanjay P K

Research Scholar, JNTU Anantpur, Assistant Professor, Department of Computer Science and Engineering, RNS Institute of Technology, Bengaluru 560098

Dr. G T Raju

Professor and Head, Department of Computer Science and Engineering, RNS Institute of Technology, Bengaluru 560098

Dr. B Eswara Reddy

Professor Department of CSE, JNTUA College of Engineering, Kalikiri

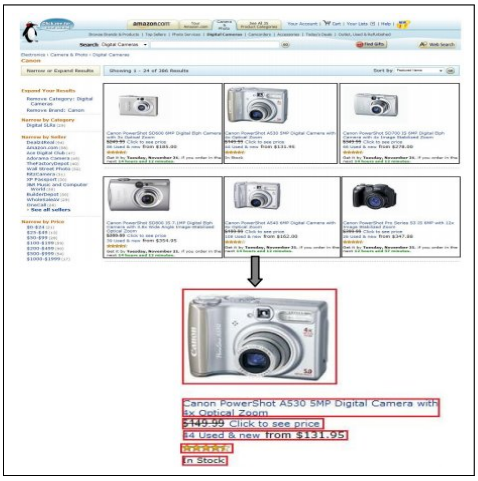
Abstract

Web object retrieval has become essential to provide vertical search facility on the Web. To achieve high quality Web object retrieval, there is a need for effective object extraction techniques and object ranking functions. In this work, an extensive survey of existing Web Object Retrieval Techniques is performed by going through more than 30 recent and relevant papers. Other techniques which are similar to Web object retrieval procedures are also described in this survey.

1 Introduction

The Web has been instrumental in providing Information Retrieval(IR) facility from the past couple of decades [1]. The central feature of Web has been to search and retrieve related documents for a given query. But, these documents also contain various objects such as people, papers, products etc. Currently, many Web users require IR at object level rather than at document level [2].

Web documents can either be Web pages or Web database records. Consider an application, which provides information about certain products. The user is provided a query form which will be filled up by the user and submitted to the Web application engine. The Web application engine uses the existing Web repositories and databases to extract the relevant answer and provides it to the user. The user might be provided with documents from which he has to search for the required information. Also, the Web databases might store old and partial data which might not be useful for the user. To overcome this bottleneck, data integration solutions can be used. By identifying object schema, information about different objects can be extracted from the existing Web databases and then, redundant information can be filtered. Finally, by performing data integration, Web object databases can be built. These Web object databases can be used



Canon PowerShot A530 5MP Digital Camera with 4x Optical Zoom

~~$149.99~~ Click to see price

44 Used & new from $131.95

In Stock

Figure 1: Web Page Having Multiple Web Objects

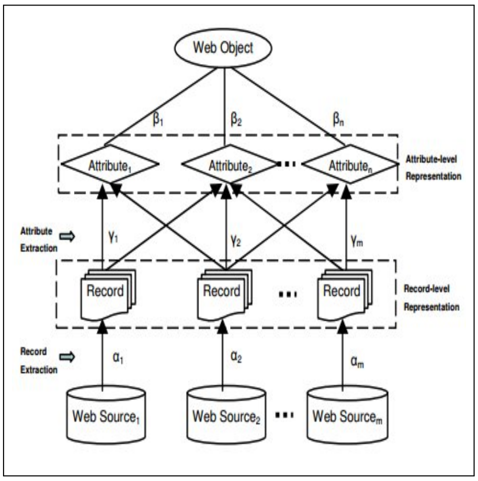


Figure 2: Web Object Extraction System

to build powerful vertical search Web applications which provide the latest and relevant information to the user [2].

To perform Web object integration, many techniques such as data record extraction [3–5], attribute value extraction [6] and object identification [7] are required. The extraction process can lead to different errors such as:

1. *Source level errors.* The quality of the information content of Web databases may not be updated or it might be redundant.

2. *Record level errors.* The data records which are extracted might have partial or unnecessary information.

3. *Attribute level error.* In some data records the required attribute values might be merged with other attribute values. For example, in some publication repository, author names are concatenated together with publication title.

Each Web object comprises of set of attributes. In Figure 1, a Web page which contains 6 data records and each record containing 6 attributes is shown. Each of the data record can be seen as Web object. Information about this particular object can be extracted from various other data sources. This collected information from different data sources can be integrated to create Web object databases. The different steps in achieving this goal are:

1. The first step is to extract information about predefined Web objects from different data sources which can be HTML documents, PDF, PS and other formats.

2. Each extracted object should be properly mapped to the predefined Web objects and then store in a single data warehouse.

3. Finally, retrieval methods should be provided to the user in-order to query stored Web objects.

The architecture of Web object extraction system is shown in Figure 2. The required data records are extracted from different Web data sources. Here, α_i indicates the accuracy of extracting the required record i from a data source. After this, the required attributes are extracted from the available records. Here, γ_i indicates the accuracy of extracting the required attribute i from the available records. From these extracted attributes, the required Web object is created. Here, β_j describes the importance of attribute j in creating the Web object.

2 Survey of Existing Techniques

2.1 Data Record and Attribute Value Retrieval

Currently, data record extraction is performed for Web pages [5, 8–11] and Web databases [3, 4]. The Web pages are divided into number of blocks and each block can be assumed as a data record. Relevance score is assigned for each block in a Web page. This style of IR is known as Passage or Block IR. The problem of assigning labels to the values of extracted attributes is addressed in [6]. Conditional random fields are used in resolving ambiguity for choosing labels. These techniques are only concerned with extracting the required data records or attributes and then ranking them. However, they do not perform Web object database creation so that, vertical search facility for the required Web objects can be provided.

2.2 XML Record Retrieval

XML documents follow a tree structure. The task is to extract certain branches of the tree which are considered as data records. Currently, no extraction process is performed on XML documents to achieve data record extraction. Many works has been proposed which deal with handling the XML query language [12], dealing with identifying the length of XML elements which helps in separating over lapped XML elements [13, 14] and test cases have been proposed which provide evaluation of effectiveness of XML retrieval [14]. Similar to XML documents, Web pages are also considered as structure documents and by utilizing this structure information, retrieval of relevant documents is performed [15–18].

2.3 Distributed Information Retrieval

Distributed information retrieval has certain similarity with Web object retrieval. The task here is to select the relevant information from different data sources and then integrating them to provide relevant result to the user. But, the techniques that are employed to perform source identification and result integration cannot be directly used for Web object retrieval. This is because even though both these problems look similar, they have different system structure which makes application of common techniques infeasible. Many data integration techniques have been proposed to achieve distributed information retrieval [19–22]. Also, many techniques have been proposed to evaluate the quality of data sources [23, 24].

2.4 Ranking of Web Objects

In [25], the Web object database is created through extraction and data integration. This work primarily focuses on ranking of Web objects. Both data records and attributes are considered as Web objects. Separate ranking functions are developed for ranking data records and attribute values.

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

The ranking function used for ranking the data records is shown in Equation 1. Here, $P(w|R_k)$ indicates the probability of generating the query term w from data record R_k, higher the probability more will be the relevance of R_k wrt w. Here, $tf(w, R_k)$ is the frequency of w in R_k, $tf(w, C)$ is the frequency of w in document collection C, $|R_k|$ is the length of document R_k, $|C|$ is the number of terms in the whole connection and λ is a smoothing parameter.

$$P(w|O_{jk}) = \lambda \frac{tf(w,O_{jk})}{|O_{jk}|} + (1-\lambda)\frac{tf(w,C_j)}{|C_j|} \qquad (2)$$

The ranking function shown in Equation 2 performs ranking of attribute values. Here, $P(w|O_{jk})$ is the probability of generating query term w from attribute value O_{jk} which is present in record R_k and C_j is the collection of all objects which contain the attribute j.

2.5 Spatial Web Object Retrieval

Currently, many Web documents are being geo-referenced and geo-tagged. For example, a user might pose a query which wants to find out all the good restaurants which serve pizza and are nearer to the users hotel. In a conventional Web object IR system, integration of Web objects are performed which might not provide indexing on the spatial location of these Web objects. Hence, to answer the above query, the conventional Web object IR system might be inefficient.

In [26], efficient IR system design is proposed to answer queries which require both information about the Web objects and their spatial locations. The spatial component of a query can be described by either a point or a rectangle. A new indexing mechanism called IR tree as shown in Figure 3 is proposed. This IR index is an extension of R tree. Finally, by using a scoring function, the ranking of spatial Web objects is performed.

2.6 Entity Retrieval

Recently, Web of Data has emerged to become one of the pillars of information sharing on the Web. It is supported by industry standards such as RDF. Basically, it is a knowledge repository in which dataset is composed of set of triples (s, p, o). Here, s is the subject, p is the predicate and o is the object. The subject contains the URI of an entity and object contains URI or literal. An entity profile e is the set of triples which has the same URI in the subject field. Every entity has a *type* which describes the given entity.

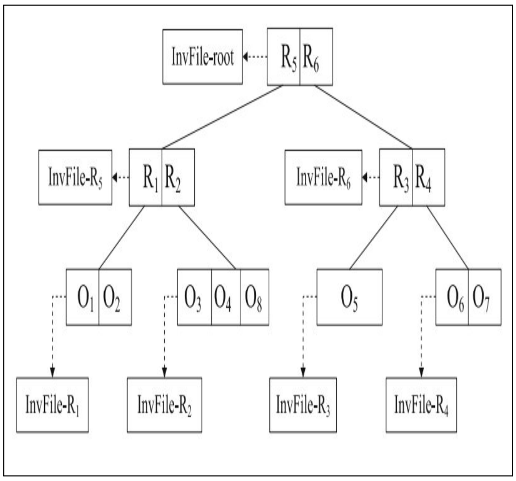


Figure 3: Spatial IR Index

The Web of Data contains numerous entities which are interlinked based on the dependency. This knowledge base helps in providing semantic search on the Web by using keyword search queries. For example, the user might provide the query *best cold medication.* To answer this query, the query processor filters the term *best* and chooses the terms *(cold, medication).* These terms are mapped to the entity domain *Medical.* The relevant entities belonging to this domain are retrieved as a result to the keyword query. This new knowledge representation provides convenience in understanding the context of the query and provides unambiguous results.

This entity retrieval has certain similarities with Web object retrieval. In Web object retrieval, objects are extracted and integrated to provide vertical search to the users. In entity retrieval, entities are already structured which can be directly used for providing vertical search on entities.

Many, entity retrieval mechanisms have been proposed [27–30]. In [31] formal model is provided to access the quality of entity retrieval systems. Coreference helps in analyzing whether two entities describe different concepts. Sometimes, the result set might contain many entities which describe the same concept. Hence, coreference aware entity retrieval was introduced in [32]. A metric called conciseness shown in Equation 3 is used to evaluate the diversity of the result set. Here, R indicates the result set.

$$Concisness(R) = \frac{number\ of\ unique\ objects \in R}{|R|} \tag{3}$$

The summary of Web object retrieval techniques is illustrated in Table 1.

3 Open Issues

There are two major open issues identified after a thorough review on existing Web object retrieval techniques:

1. To perform Web object data integration, effective object extraction methods are required to resolve ambiguity. There is a wide scope in improving the effectiveness of the object extraction techniques.

2. Currently, only a single ranking function is available for ranking the Web objects. There is a need to design better ranking functions to achieve high effectiveness in Web object retrieval.

Techniques	Features	Merits	Limitations
Data Record and attribute value retrieval [3–5, 8–11]	Performs extraction of data records and attribute values in order to achieve object integration	Effective and efficient Web object retrieval	Does not perform ranking of Web objects
XML Record Retrieval [12–14]	Performs record retrieval in XML documents by utilizing XML tags	Helps in extracting relevant part of the XML documents in-order to provide relevant information	Techniques are not suitable for other non XML documents
Distributed Information Retrieval [19–24]	Techniques achieve retrieval of relevant information in a distributed environment	Scalable information retrieval	Not suitable to be used directly for Web object retrieval
Ranking of Web objects [25]	Performs ranking of retrieved Web objects for a given query	Effectiveness in terms of user relevant results	Scoring function can be improved further
Spatial Web object retrieval [26]	Uses novel IR index to perform spatial Web object retrieval	Efficient Spatial Web object retrieval	Only suitable for spatial objects
Entity retrieval [27–32]	Achieves effective IR on knowledge databases	Provides user with unambiguous result	Techniques are only suitable for knowledge databases

Table 1: Summary of Web object retrieval techniques

4 Conclusion

In this work, the problem of Web object retrieval is described along with its necessity. Survey of existing Web object retrieval techniques and other similar techniques is presented. The existing open issues which provide an opportunity to conduct future research in this area are detailed. There is still an ample opportunity to design effective Web object retrieval techniques to provide powerful vertical search engines.

References

[1] Ricardo Baeza-Yates and Berthier Ribiero-Neto. *Modern Information Retrieval* Addison-Wesley publishers, 1999.

[2] Zaiqing Nie, Yuanzhi, 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.

[3] K.Lerman, L.Getoor,S.Minton and C.A.Knoblock. *Using the Structure of Web sites for automatic segmentation of tables* In ACM SIGMOD Conference (SIGMOD), 2004.

[4] J.Wang and F.H.Lochovsky. *Data extraction and label assignment for Web databases.* In World Wide Web conference (WWW), 2003.

[5] Bing Liu, Robert Grossman and Yanhong Zhai. *Mining Data Records in Web pages.* ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2003.

[6] 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.

[7] S.Tejada, C.A.Knoblock and S.Minton. *Learning domain-independent string transformation weights for high accuracy object identification.* In knowledge Discovery and Data Mining (KDD), 2002.

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

[9] Deng Cai, Shipeng Yu, Ji-Rong Wen and Wei-Ying Ma. *Block-based Web search.* In processing of SIGIR, 2004.

[10] J.P Callan. *Passag-Level Evidence in Document Retrieval.* In Processing of SIGIR, 1994.

[11] M.Kaszkiel and J. Zobel. *Passage Retrieval Revisited.* In Proceedings of SIGIR, 1997.

[12] Norbert fauhr and Kai Grojohann. XIRQL. *A Query Language for Information Retrieval in XML documents.* In proceedings of the SIGIR, 2001.

[13] Jaap kamps, Maarten de Rijke and Borkur Sigurbjornsson. *Length normalization in XML retrieval.* In proceedings of the SIGIR, 2004.

[14] Charles L.A Clarke. *Controlling Overlap in Content oriented XML Retrieval.* In proceedings of the SIGIR, 2005.

[15] Paul Ogilvie and Jamie Callan. *Combining Document Representation for Known item search.* In proceedings of SIGIR, 2003.

[16] Thijs Westerveld, Wessel Kraaij and Djoerd Hiemstra. *retrieving Web Pages using Content, Links, URLs and Anchors.*

In The Tenth Text REtrieval Conference (TREC2001), 2001.

[17] Nick Craswell, David Hawking and Trystan Upstill. *TREC 12Web and Interactive Tracks at CSIRO.* In The Twelth Text REtrieval Conference(TREC 2003), 2004.

[18] Abdur Chowdhury, Mohammed Aljlay, Eric Jensen, Steve Beitzel, David Grossman and Ophir Frieder. *Linear Combination Based on Documents Structure and varied Stemming for Arabic retrieval.* In The Eleventh Text REtrieval Conference(TREC2002). 2003.

[19] J.P Callan. *Distributed information retrieval.* In Advances in Information Retrieval: Recenr research from the center for Intelligent Information Retrieval, edited by W.Bruce Croft.Kluwer Academic Publisher, pp. 127-150, 2000.

[20] L.Gravano and H.Garcia-Molina *Generalizing gloss to vector-space databases and broker hierarchies.* In Proceedings of the International Conference on Very Large Databases(VLDB), 1995.

[21] M.Meng, K.Liu, C.Yu, W.Wu and N.Rishe. *Estimating the usefulness of search engines.* In ICDE Conference, 1999.

[22] J.Xu and J.Callan. *Effective retrieval with distributed collections* In Proceedings of SIGIR, 1998.

[23] Amihai Motro and Igor Rakov. *Estimating the quality of databases.* In Proceedings of the 3rd International Conference on Flexible Query Answering(FQAS), Roskilde, Denmark, May 1998. Springer Verlag.

[24] Felix Naumann and Rolker Claudia. *Assessment Methods for Information Quality Criteria.* In Proceedings of the International Conference on Information Quality (IQ), Cambridge, MA, 2000.

[25] Zaiqing Nie, Yunxiao Ma, Shuming Shi, Ji-Rong Wen, Wei-Ying Ma *Web Object Retrieval.* WWW 2007, May8-12, 2007, Banff, Alberta, Canada.

[26] Dingming Wu, Gao Cong, Christian S.Jensen, *A frame work for efficient spatial Web object retrieval*, The VLDB journal(2012) 21:797-822, DOI 10.1007/s00778-012-0271-0.

[27] Jiwei Ding, Wentao Ding. Wei Hu, Yuzhong Qu, *An EBMC-based approach to selecting types for entity filtering*, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, P.88-94, January 25-30, 2015, Austin Texas.

[28] Besnik Fetahu, Ujwal Gadiraju, Stefen Dietze, *Improving Entity Retrieval on Structured Data*, Proceedings of the 14th International Conference on The Semantic Web-ISWC 2015, October 11-15, 2015.

[29] Nikita Zhiltsov, Alexander Kotov, Fedor Nikolaev, *Fielded Sequential Dependence Model for Ad-Hoc Entity Retrieval in the Web of Data*, Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, August 09-13, 2015, Santiago, Chile.

[30] Aliaksandr Talaika, Joanna Biega, Antoine Amarilli, *Fabin M.Suchanek, IBEX:Harvesting Entities from the Web Using Identifiers*, Proceedings of the 18th International Workshop on Web and Databases, May 31-june 04, 2015, Melbourne, VIC, Australia.

[31] J.Pound, P. Mika and H.Zaragoza. *Adhoc object retrieval in the Web of data.* In

WWW10, Pages 771-780, New York, NY, USA, 2010.ACM.

[32] Jeffrey Dalton, Peter Mika and Roi Blanco. *Coreference Aware Web Object Retrieval.* CIKM11, October 24-28,2011, Glasgow, Scotland, UK.