Query-Related Data Extraction of Hidden Web Documents

Y.L. Hedley, M. Younas, A. James
School of Mathematical and Information Sciences
Coventry University, Coventry CV1 5FB, UK
{y.hedley, m.younas, a.james}@coventry.ac.uk

M. Sanderson
Department of Information Studies
University of Sheffield, Sheffield, S1 4DP, UK
m.sanderson@sheffield.ac.uk

ABSTRACT

The larger amount of information on the Web is stored in document databases and is not indexed by general-purpose search engines (i.e., Google and Yahoo). Such information is dynamically generated through querying databases — which are referred to as Hidden Web databases. Documents returned in response to a user query are typically presented using template-generated Web pages. This paper proposes a novel approach that identifies Web page templates by analysing the textual contents and the adjacent tag structures of a document in order to extract query-related data. Preliminary results demonstrate that our approach effectively detects templates and retrieves data with high recall and precision.

Categories and Subject Descriptors: H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services.*

General Terms: Performance, Experimentation.

Keywords: Hidden Web Databases, Data Extraction.

1. INTRODUCTION

Hidden Web databases [3] maintain a collection of documents (i.e., archives, manuals or news articles) and dynamically generate a document list in response to user queries. This is beyond the indexing capability of search engines, which crawl Web pages through hyperlinks.

Recent research discovers the content of a database through sampling its documents [1, 4, 8]. The terms and statistical information gathered from sample documents is referred to as 'Language Models' [1], 'Textual Models' [4, 8] or 'Centroids' [5]. This is then used for the purpose of database selection [1, 4, 8] or database categorisation [5]. However, a number of terms extracted are often found in templates used for description or navigation purposes, thus they are irrelevant to a user query. For instance, Language Models [1] consists of terms (such as 'Author' and 'Home') with high frequencies that are not relevant to the database content when sampling documents from Combined Health Information Database (CHID). [1] therefore proposes the use of additional stop-word lists to eliminate irrelevant terms, but maintains that such a technique can be difficult to apply in practice. Textual Models [4, 8] contain additional topic terms through sampling Web databases. However, their techniques do not consider the elimination of terms contained in templates.

Approximate string matching techniques are adopted by [6] to extract information from Web pages. This approach is limited to the similarities and differences of textual contents. [2] extracts dynamically generated objects from Web pages by analysing tags and textual contents based on tree-like structures. This requires extra computation for converting and analysing Web pages in tree-like structures.

In this paper we propose a new approach that identifies the portions of Web documents that are relevant to a user query by analysing textual contents and their adjacent tag structures. This approach considers a document to be a list of tags and texts (i.e., segments between tags). Each text segment is then represented by its textual content along with the adjacent tag segments. Adjacent tag structures of a text segment describe how the text is presented in a Web document or how the text relates to its neighbouring text. This representation is then used to detect templates. Preliminary results show that our technique provides an effective mechanism to extract query-related data from Hidden Web databases.

2. QUERY-RELATED DATA EXTRACTION

This section focuses on the extraction of data relevant to a user query from a Hidden Web document - which we refer to as *query-related data*. The proposed approach is presented in three phases. Firstly, how document contents are processed and represented is described. The second phase provides the mechanism to identify Web page templates. The final phase demonstrates the process that determines the similarities between text segments of different documents.

2.1 Document Content Representation

This phase converts documents retrieved from databases into a list of tag segments and text segments. Tag segments include starting tags, ending tags or single tags. Text segments are texts that reside between tag segments. Each text segment can further be identified by its adjacent tag segments. Adjacent tag segments of a text segment are defined as the tags that are located before and after the text segment. A text segment is then defined as follows.

TextSegment = {text, tag, tag}

For example, consider the following segments in a document.

<TABLE>…</TABLE>
Links:
<A> documentation </A>
<A> tutorials </A> …

Text segment 'Links:' is therefore represented as *{"Links:", </TABLE>, <A>}*.

2.2 Template Detection

Documents retrieved from Hidden Web databases are often presented using one or more templates. The mechanism to detect

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

templates is described as follows: (i) Text segments of documents are analysed based on textual contents and their adjacent tag segments (ii) An initial template is identified by examining the first two sample documents (iii) The template is then generated if matched text segments along with their adjacent tag segments are found from both documents (iv) Subsequent documents retrieved are compared with the template generated. Text segments that are not found in the template are extracted for each document to be further processed (v) When no matches are found from the existing template, document contents are extracted for the generation of future templates. The process is repeated until the required number of documents is sampled. This generates a set of templates - each with a list of documents that contain potential query-related data.

2.3 Text Similarity Computation

The text segments extracted from documents are further analysed. This process identifies document contents that have not been found in the templates generated from the initial sample documents.

In this phase, the text of a segment is represented as a vector of terms with weights. A term weight is obtained from the frequencies of the term that appears in the segment. Cosine similarity [7] is computed for the text segments of different documents that are generated from the same template in order to determine their similarities. The computation of similarity for two text segments is given as follows.

$$COSINE(Text_i, Text_j) = \sum_{k=1}^{t} (TERM_{ik} \cdot TERM_{jk}) \Bigg/ \sqrt{\sum_{k=1}^{t} (TERM_{ik})^2} \cdot \sqrt{\sum_{k=1}^{t} (TERM_{jk})^2}$$

where $TERM_{ik}$ is the weight of term k in the first text, and $TERM_{jk}$ is its weight in the second text. The similarities are computed for text segments with identical adjacent tag segments only. Two segments are considered to be similar if the similarity exceeds a threshold value. Such a threshold value is determined experimentally. This process extracts text segments that are significantly different from textual content and tag structures contained in templates.

3. PRELIMINARY RESULTS

Experiments are carried out on 8 Web databases that provide user manuals, archives and news articles. Each database is randomly sampled to retrieve a subset of documents. 10 documents are retrieved from each database and a total of 80 documents are analysed. Sample documents are manually examined to obtain the number of templates used in each database and that of templates which have been detected. Terms extracted using the proposed technique are also manually compared with the document from which the terms originate. Recall and precision techniques (of information retrieval systems [7]) are modified in order to measure the accuracy of query-related data extraction. The modified recall is given by the ratio of the number of relevant terms retrieved over the total number of relevant terms contained in a document. The modified precision is defined as the ratio of the number of relevant terms retrieved over the total number of terms retrieved from a document. The results shown in Table 1 demonstrate that our approach: (i) accurately detects the number of templates used by each database (ii) extracts query-related data with high recall and precision. The overall accuracy for all sample database documents is 96.6% and 94.7% in terms of recall and precision respectively.

Table 1. The performance of query-related data extraction

Document databases	Number of templates		Query-related data extraction	
	Used	Detected	Average recall	Average precision
help-site.com	4	4	98.0%	97.9%
devx.com	3	3	94.7%	93.3%
simplethebest.net	1	1	91.6%	95.9%
techweb.com	1	1	98.8%	82.0%
wired.com	3	3	99.9%	98.6%
reed-electronic.com	1	1	97.2%	99.0%
itpapers.zdnet.com	1	1	98.7%	93.1%
chid.nih.gov	1	1	94.1%	99.9%

4. CONCLUSIONS AND FUTURE WORK

Recent research demonstrates that the contents of databases can be represented by terms and frequencies retrieved from randomly sampled documents. However, Hidden Web databases dynamically generate contents (such as archives and news articles) using templates. Therefore, we propose a new technique that detects templates based on textual contents and their adjacent tag structures. The aim is to extract query-related data in order to obtain terms and frequencies with a higher degree of accuracy. Our technique examines text segments along with their adjacent tag structures rather than analysing document contents in a tree-like structure as in [2]. This provides an effective mechanism for template detection. The experiments demonstrate that this technique has attained recall and precision of high accuracy.

Future work includes the experiments of the proposed technique on a larger sample of Hidden Web documents. Moreover, our technique will be applied to obtain a representative set of terms and frequencies from databases for their categorisation.

5. REFERENCES

[1] Callan, J., and Connell, M. *Query-Based Sampling of Text Databases*. ACM Transactions on Information Systems, Vol. 19, No. 2, 2001, 97-130.

[2] Caverlee, J., Buttler, D. and Liu, L. *Discovering Objects in Dynamically-Generated Web Pages*. Technical report, Georgia Institute of Technology, 2003.

[3] Gravano, L., Ipeirotis, P. G. and Sahami, M. *QProber: A System for Automatic Classification of Hidden-Web Databases*. ACM Transactions on Information Systems (TOIS), Vol. 21, No. 1, 2003.

[4] Lin, K.I. and Chen, H. *Automatic Information Discovery from the Invisible Web*. International Conference on Information Technology: Coding and Computing, 2002.

[5] Meng, W., Wang, W., Sun, H. and Yu, C. *Concept Hierarchy Based Text Database Categorization*. International Journal on Knowledge and Information Systems, Vol. 4, No. 2, 2002, 132-150.

[6] Rahardjo, B. and Yap, R. *Automatic Information Extraction from Web Pages*, SIGIR, 2001, 430-431.

[7] Salton, G. and McGill, M. *Introduction to Modern Information Retrieval*. New York, McCraw-Hill, 1983.

[8] Sugiura, A. and Etzioni, O. *Query Routing for Web Search Engines: Architecture and Experiment*. 9th WWW Conference, 2000.