KNOWLEDGE DISCOVERY FROM THE WEB

Maryam Hazman
Central Lab for Agricultural Expert Systems Agricultural Research Center, Ministry of Agriculture and Land Reclamation, Giza, Egypt

Samhaa R. El-Beltagy
Computer Science Department, Faculty of Computers and Information, Cairo University, Giza, Egypt

Ahmed Rafea
Computer Science Department, American university, Cairo, Egypt

Salwa El-Gamal
Computer Science Department, Faculty of Computers and Information, Cairo University Giza, Egypt

Keywords: knowledge discovery, indexing, web content mining.

Abstract: The World Wide Web is a rich resource of information and knowledge. Within this resource, finding relevant answers to some given question is often a time consuming activity for a user. In the presented work we construct a web mining technique that can extract information from the web and create knowledge from it. The extracted knowledge can be used to respond more intelligently to user requests within the diagnosis domain. Our system has three main phases namely: a categorization phase, an indexing phase, and search a phase. The categorization phase is concerned with extracting important words/phrases from web pages then generating the categories included in them. The indexing phase is concerned with indexing web page sections. While the search phase interacts with the user in order to find relevant answers to their questions. The system was tested using a training web pages set for the categorization phase. Work in the indexing and search phase is still in going.

1 INTRODUCTION

The World Wide Web is a huge collection of texts, which is constantly growing. This amount of text is a valuable resource of information and knowledge. Finding useful information in this resource is not an easy task. People want to extract useful information from these texts quickly and at a low cost (Loh *et al*., 2000). They prefer reaching a certain paragraph, which is of concern to them instead of reading an entire document. In addition, when searching for an item of interest using a traditional search engine, numerous results are returned which requires the user to manually try to filter through these results. This places an overhead on the user in terms of item and effort. Research in Web mining is moving the World Wide Web towards a more useful environment in which users can quickly and easily find information they need (Scime, 2004).

This paper addresses the particular problem of trying to find a certain section that is of concern to a user. The goal of the research described here is to develop a Web mining system that can be used for answering a user request within the diagnosis domain. To perform this goal, the system automatically discovers categories in a set of web documents within some predefine domain. The system then uses these categories to classify various sections from other web pages within the same domain. If a section belongs to a diagnosis category, it will be indexed and stored in a database table. When a user submits a query in the form of some observations, the system will look for them in the index table. The answer is provided in the form of a disorder category with a link to its source section in

Hazman M., R. El-Beltagy S., Rafea A. and El-Gamal S. (2005).
KNOWLEDGE DISCOVERY FROM THE WEB.
In *Proceedings of the Seventh International Conference on Enterprise Information Systems*, pages 303-308
DOI: 10.5220/0002547103030308

the web page. To achieve this goal, we build a system with three main phases: a categorization phase, an indexing phase, and a search phase. The categorization phase extracts categories included in a training web pages set and classifies these categories. The indexing phase assigns a category for each section in a web page section if possible and indexes the diagnostic sections. The search phase interacts with the user to search for relevant answers to their question.

The rest of this paper is organized as follows: section 2 presents related work. Section 3 describes the architecture of the proposed system. Sections 4, 5, 6, 7, and 8 present the categories extraction module, categories module and heuristic rules component, the indexing module, repository component and knowledge finder with the user interface respectively. Finally section 9 provides the conclusion for the presented research.

2 RELATED WORK

Zaiane defines Web mining as the extraction of interesting and potentially useful patterns and implicit information from artifacts or activities related to the World Wide Web (Zaiane, 1999). Web mining research is at the cross roads of research from several research communities, such as database, information retrieval, and Artificial Intelligence, especially the sub-areas of machine learning and natural language processing. However, there is confusion when comparing efforts from different points of view (Kosala and Blockeel, 2000). Today the most recognized categories of the Web mining fall into three areas of interest based on the type of Web data to be mined: web content mining, web structure mining, and web usage mining (Kosala and Blockeel, 2000), (Doherty, 2000), (Borges and Levene, 1999), (Madria *et al*., 1999), (Pal *et al.*, 2002). In practice, the three Web mining tasks could be used in isolation or combined in an application (Kosala and Blockeel, 2000). Web structure mining is the process of inferring knowledge from World Wide Web organization and links between references and referents in the Web. Web usage mining mines the secondary data derived from the interactions of the users while interacting with the web (Kosala and Blockeel, 2000). Web content mining is concerned with the discovery of new information and knowledge from web based data, documents, and pages (Hsu, 2002). It is mainly based on research in information retrieval and text mining, such as information extraction, text classification and clustering, and information visualization. However, it also includes some new applications, e.g., Web resource discovery (Chen and Chau 2004). Our work on knowledge discovery from Web is related to web content mining.

KPS is an information mining algorithm. It employs keywords, patterns and/or samples to extract information from semi-structured textual Web pages to mine the desired information (Guan and Wong, 1999). El-Beltagy *et al.* (2004) present a model for information extraction and intelligent search. It automates augmenting segments of organizational documents that cover similar concepts within a known domain with metadata. The model uses dynamically acquired background domain knowledge in order to construct the documents categories. Liu *et al*. (2003) propose a set of effective techniques to perform the task of mining and organizing topic-specific knowledge on the Web. They find and compile topic specific knowledge (concepts and definitions) on the Web. Loh *et al*. (2000) present an approach for knowledge discovery in texts extracted from the web. Instead of analyzing words or attribute values, the approach is based on concepts, which are extracted from texts to be used as characteristics in the mining process. Statistical techniques are applied on concepts in order to find interesting patterns in concept distributions or associations. For identifying concepts in texts, a categorization algorithm is used associated to a previous classification task for concept definitions (Loh *et al*., 2000). In (Xu *et al*., 2003) a research support system framework for web data mining is presented. This framework is designed for identifying, extracting, filtering and analyzing data from web resources. It combines web retrieval and data mining techniques together to provide an efficient infrastructure to support web data mining for research (Xu *et al*., 2003).

3 ARCHITECTURE

Recall from section one above that our proposed system operates in three phases: a categorization phase, an indexing phase, and a search phase. Figure 1 depicts the architecture of the proposed system in terms of the main components and their interactions. The categorization phase is concerned with using the structure of some input documents (training web pages) in order to determine categories and sub categories that will generalize across some given domain. This phase includes the categories extractor, categorizer, and heuristic rules applier components. The categories extractor parses the training web pages set content to extract important words/phrases that represent categories included in it, and generates their corresponding knowledge XML file. Then, the

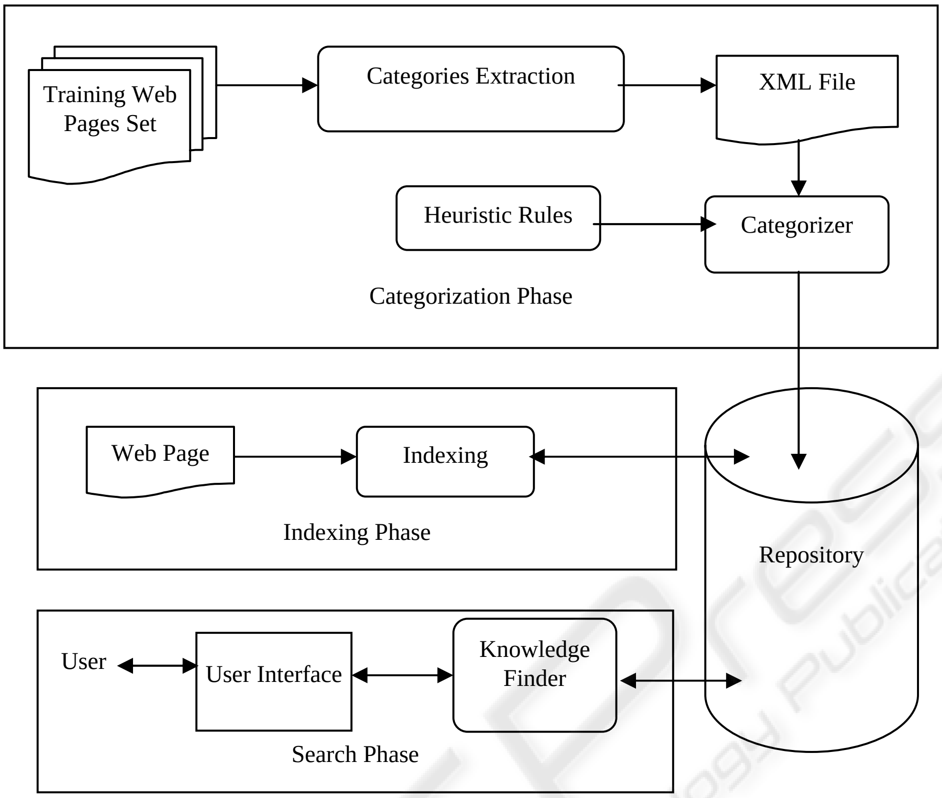

Figure 1: The proposed system architecture.

categorizer automatically generates the main categories, subcategories, and sub-subcategories from XML using a set of hypothesis rules. The indexing phase includes the indexing component, which is responsible for indexing web page sections. The search phase looks for the user query in the repository. It includes the knowledge finder and the user interface. The three phases are linked together through the repository component.

The following sections describe the goals and functionally of the above components in more details.

4 THE CATEGORY EXTRACTOR

The objective of this component is to extract important words/phrases in a training web pages set. Each set of words or phrases would then represent the category of its section, where html heading tags (e.g., <h1>,.,<h4>) are used to define their importance. Relationships between categories and subcategories can also be deduced using heading information (e.g., <h2> tag is a child for <h1> tag). The heading information represents the dependency between the sections in a web page. For example given a set of documents in the agricultural domain, the Diseases section can be followed by detailed sections about specific disease categories such as Fungal disease, which in turn will be followed by specific Fungal disease instances like Wilt Root Rot. So "Diseases" is the main category, Fungal is its subcategory, and Wilt Root Rot is a subcategory of Fungal. In other word, "Diseases" is the parent of Fungal disease, which is a parent of Wilt Root Rot. Important words/phrases extracted must be converted into a more structured representation that can be used to classify the extracted categories. To do that, the categories extractor parses the training web pages set content and generates their corresponding knowledge XML files. This procedure is performed offline once, and should only be repeated if the content of a Web page has a new category missing in the categories databases.

Although some of the extracted important words/phrases are marked as heading html tags, they do not in fact represent actual categories. For this reason, we have identified the following rules to determine if an important phrase can be safely ignored: -

```
<Root>
   <Category>
        <Name> Disease </Name>
        <HeadingLevel> 1 </HeadingLevel>
  </Category >
  <Category >
        <Name> Fungal </Name >
        <HeadingLevel> 2 </HeadingLevel>
  </Category >
  <Category >
        <Name> Wilt Root Rot </Name>
        <HeadingLevel > 3 </HeadingLevel>
  </Category >
  <Category >
        <Name> Nutrition Deficient </Name>
        <HeadingLevel> 1 </HeadingLevel>
  </Category >
  <Category >
        <Name> Nitrogen Deficient </Name>
        <HeadingLevel> 1 </HeadingLevel>
  </Category >
  <Category >
        <Name > Potassium Deficient </Name>
        <HeadingLevel> 1 </HeadingLevel>
  </Category >
</Root>
```

Figure 2: The structure of the generated XML file from the category extractor component

- Contains the word introduction.
- Contains the word return.
- Contains the word note.
- Contains the word advice.
- Contains the word recommendation.
- Contains the word guidance.
- Contains the word warning.
- Contains the publication data (date, author, publisher).
- Is too lengthy a phrase (we use 8 words as the upper limit for a useful phrase).

Figure 2 represents an example of a XML file structure generated from this component.

5 THE CATEGORIZER COMPONENT

The objective of the categorizer is to automatically generate the main categories, sub-categories, and sub-subcategories. It parses all the XML nodes, and for each node performs stopword removal and word stemming, which are standard operations in information retrieval. Stopwords are words that occur frequently in documents and have little informational meaning. The process of stemming finds the root of a word by removing its suffix and prefix. We are used the lexical analyser for inflected Arabic words described in (Rafea and Shaalan, 1993) to get a word's stem.

After obtaining the list of stems for each important phrase that represents a category, the important phrase along with its stem list is stored in the database. The relationship between any given category and other categories (child, parent, grandparent) is also stored. Following this step, we apply our mining algorithm, which is basically a set of heuristic rules in order to obtain a set of categories and subcategories that we can later use. These rules are as follows (written in first order predicate logic): -

- The main category is a category that has no parent in all the training set pages or for which the number of times it appears to have no parent, exceeds the number of times that it appears to have a parent.
  $\text{Main Category}(X) \rightarrow \text{Category}(X) \wedge \neg\exists Y\ \text{parent}(Y,X)$
  $\text{Main Category}(X) \rightarrow \text{Category}(X) \wedge \text{Count}(\text{parent}(\text{NoParent},X),N) \wedge \text{Count}(\text{parent}(Y,X),M) \wedge N >= M$
- A subcategory is a category that has a parent in all training set pages or for which the number of times it appears to have parent, exceeds the number of times that it appears to have no parent.

$\text{SubCategory}(X) \rightarrow \text{Category}(X) \wedge$
$\exists Y\ \text{parent}(Y,X)$
$\text{SubCategory}(X) \rightarrow \text{Category}(X) \wedge$
$\text{Count}(\text{parent}(\text{NoParent},X),N) \wedge$
$\text{Count}(\text{parent}(Y,X),M) \wedge$
$M > N$

- A sub-subcategory is a category that has a grandparent in all the training set pages.
  $\text{SubSubCategory}\ (X) \rightarrow \text{Category}(X) \wedge$
  $\exists Z\ \text{grandparent}(Z,X)$
- Two categories are the same if they have the same stemmed word set.
  $\text{SameCategory}\ (X,Y) \rightarrow \text{Category}(X) \wedge$
  $\text{Category}(Y) \wedge$
  $(\text{list}(X) = \text{list}(Y))$

6 THE INDEXING COMPONENT

The main goal of the indexing component is to index the diagnostic sections for web pages in some given domain. The domain we have worked on is the agricultural domain. The indexing component assigns a category to each section in a web page if possible. It extracts a section's important words/phrases then removes the stopwords and gets their stem for determining the corresponding category in the database. If a category does not exist in the repository database then the categorizer component is called. The categorizer determines if there is a need for adding this category in the repository as a new category or not.

Our index is not like other search engine indexes; its purpose is to answer a query about a problem in the diagnostic domain. For example, a user can submit a query in the form of observations about a plant to the system and get the reasoning of it as an answer to his/her query. To build such an index, we need taxonomy and some meta knowledge. In our implementation, we have defined a domain-specific taxonomy for the agricultural diagnostic domain in a database table. At the end of the indexing stage, each web page is characterized by a set of categories, which are part of it.

Initially the diagnostic categories were simply categorized by a vector of stemmed words and a link to the source section from which they were extracted. However, this simple representation was found to suffer from a major drawback as relationships between words and concepts to which they are associated was lacking. As a result, submitting a query stating that "The color of leaves is red, and fruits have spots" would match with a category item which has the sentence "The fruits are red. Also, the leaves are spotted", even though it is in fact irrelevant to the submitted query.

To resolve this problem, we have extended the diagnostic category representation such that each word within the vector of stemmed words, was associated with its related concept.

7 THE REPOSITORY

The purpose of the repository is to enable the system to store the extracted knowledge for the purpose of searching it. The repository contains the full categories of every web page. Within the repository, each page is divided into sections which are stored along with their categories, and the vector characteristic their content. For diagnostic section its index is stored.

The following seven tables are used within the repository:-

1. SecIndex: stores a section's important stemmed words as a vector of words with their semantic concept.
2. SectionId: stores information about each section, such as its category, original URL, title, crop name, ..etc.
3. MainCat: stores a list of main categories.
4. SubCat: stores sub categories, linked with their main category using the mainCatId.
5. Cat: stores category items, linked with their subcategory and main category using the SubCatId and mainCatId.
6. CategoryT: stores important words/phrases of a section, as well as its parent, and its ancestor as a list of words. This table is mined to classify categories.
7. IndexWord: keeps the domain specific words needed for indexing a section.

8 KNOWLEDGE FINDER AND THR USER INTERFACE

A user can submit a query to the system via the user interface component, the user interface component then sends the user query to the knowledge finder. The knowledge finder parses the query performing stopwords removal and word stemming on it. It then, formulates and prepares an SQL query, which is sent to the repository database in order to find sections containing information relevant to the user query.

When the result is more than one section, the various sections are checked. If these sections belong to the same item category the category is returned to the user. Otherwise, it is deduced that these sections belong to different item categories, and further questions are presented to the user in order to confirm one of them. If there is a picture associated with any of these sections, it is displayed to get further confirmation. If the data entered by the user is still not enough to confirm or rule out a category, suspected categories are presented to the user with links to their original section as a reference to the user.

9 CONCLUSION

The objective of our research is to help Web users to quickly and easily find an answer to some given diagnostic question they have from specific section(s) in some given document set. To achieve this goal, we have constructed a web mining technique that can extract information from the web and create knowledge from it. Our system has been built in the agricultural domain to extract information from its related web pages, and to index the diagnostic sections in it. The constructed index is used for finding relevant knowledge to answer a user query.

Our system has three main phases: the categorization phase, the indexing phase, and the search phase. The categorization phase has been tested on a training web pages set, which is a collection of extension documents. It automatically generated 100 main categories, 145 sub categories, and 127 sub-subcategory items. These categories are used by the indexing component to assign for each section in an input web page, a category if possible. The indexing and search phases are still under construction. Also, there are still some problems must need to be solved like inheritance from more than one category, and synonymous words used in different web pages content.

REFERENCES

Borges, J. and Levene, M., 1999. Data mining of user navigation patterns, In *Web Usage Analysis and User Profiling*, vol. 1836, pp. 92-111.

Chen, H. and Chau, M., 2004. Web Mining: Machine Learning for Web Applications. In *the Annual Review of Information Science and Technology*, vol. 38, pp. 289-329.

Doherty, P., 2000. Web Mining - The E-Tailer's Holy Grail. In *DM Direct*.

El-Beltagy, S. R., Rafea, A. and Abdelhamid, Y., 2004. Using Dynamically Acquired Background Knowledge For Information Extraction And Intelligent Search. In M. Mohammadian, (Ed.) *Intelligent Agents for Data Mining and Information Retrieval*, Idea Group Publishing, Hershey, PA, USA, pp. 196-207.

Guan, T. and Wong, K., 1999. KPS: a Web information mining algorithm. In *Proceedings 8th Int. World Wide Web Conf.*, Canada, pp. 417-429.

Hsu, J., 2002. Web Mining: A Survey of World Wide Web Data Mining Research and Applications. In *Decision Sciences Institute Annual Meeting Proceedings*, PP. 753-758.

Kosala, R. and Blockeel, H., 2000. Web Mining Research: A Survey. In *SIGKDD Explorations*, vol. 2, no. 1,pp 1-15.

Liu, B., Chin, Ch. W. and Ng, H. T., 2003. Mining Topic-Specific Concepts and Definitions on the Web, In *Proceedings of the twelfth international World Wide Web conference (WWW-2003)*, Budapest, Hungry, pp. 20-24.

Loh, S., Wives, L. K. and de Oliveira, J. P. M., 2000. Concept-Based Knowledge Discovery. In *Texts Extracted from the Web SIGKDD Explorations*, vol. 2, no. 1, pp. 29-39.

Madria, S.K., Bhowmick, S.S., Ng, W.K. and Lim, E.P., 1999. Research issues in web data mining. in *Proceedings 1st International Conf. On Data Warehousing and Knowledge Discovery Florence Italy*, PP. 303-312.

Pal, S., Talwar, V., and Mitra, P., 2002. Web Mining in *Soft Computing Framework: Relevance, State of the Art and Future Directions. IEEE Trans. on Neural Networks*, 13(5):1163 -1177, 2002.

Rafea, A. and Shaalan, K.,1993. Lexical Analysis of An Inflected Arabic Word Using Exhaustive Search of an Augmented Transition Network, In Software Practice & Experience, vol. 23, no. 6, pp. 567-588.

Scime, A., 2004. Guest Editor's Introduction: Special Issue on Web Content Mining. In *Journal of Intelligent Information Systems*, vol. 22, no. 3, pp. 211-213.

Xu, J., Huang, Y. and Madey, G., 2003. A Research Support System Framework for Web Data mining Research. In Workshop on Applications, Products and Services of Web-based Support Systems at the Joint International Conference on Web Intelligence (2003 IEEE/WIC) and Intelligent Agent Technology, Halifax, Canada, October 2003, 37-41.

Zaiane, O. R., 1999. Resource and Knowledge Discovery from the Internet and Multimedia Repositories, Ph.D. thesis, Simon Fraser University.