A Framework for Automatic Classification of e-Business Web Content

Udomsit Sukakanya and Kriengkrai Porkaew

School of Information Technology
King Mongkut's University of Technology Thonburi
Thailand
Email: udomsit@sit.kmutt.ac.th, porkaew@sit.kmutt.ac.th

Abstract

Classifying the specific e-Business Web content in search results is presently done manually by human who visit a website and then define under which topic headings the web page or site belongs. This is a tedious and expensive process done by many portal websites. Unfortunately, it has become almost impossible for a casual user to look for specific e-Business information without getting lost among huge amounts of mixed data. In particular, retrieval failures arise because of the ambiguity of natural language. Queries are at best imperfect representations of the user's information needs and words chosen by the author are imperfect representations of the information contained in the document. It should come as no surprise that matching query words to document words, which is the heart of any information retrieval system, yields a very imperfect result. Moreover, the distributed nature of the WWW adds new problems to old: documents may be duplicated many times at many different sites; Web pages are added at alarming rates, creating an extremely dynamic information environment; the quality of information contained in Web pages varies greatly; Web pages are deleted or moved frequently, leaving behind dangling references.

This paper proposes a framework and a system implementation for automatic e-Business Web content classification in search results, which tries to fill the gaps mentioned above by using present research techniques including study of human retrieval behavior and other information placed inside the html code itself. In order to evaluate the system, two test sets (offline and online) were taken under consideration. For offline testing, we used 7 e-Business Web collection groups from CMU World Wide Knowledge Base, 1250 Web pages for training and 2000 Web pages for testing. For online testing, we used Web collection from the results in search engine for testing. Both of the results show that the average system performance is about 85%.

1 Introduction

Software that enhances the performance of search engines, text mining, ontology, summarization, content classification and taxonomy construction has had dramatic impact on lessening the burden of the knowledge worker [1]. Part of the user dissatisfaction stems from information management solutions in organizations that have developed organically over a period of years. Most organizations do not have centralized, strategically developed
information architecture [2]. Under such architecture, multiple layers of information management functionality are orchestrated and coordinated as a service, capable of being deployed against any number of content sources and repositories. Well defined information architecture is the antithesis of approaches that link search and management to individual repositories [3], causing user frustration with jumping from one research environment to another. In the recent Delphi Research survey [4], 82% of respondents reported that they do not have access to a centralized single point of search and management across information sources.

A well-thought-out, orchestrated approach to content tracking, categorization, search, and retrieval can give the illusion of content integration. Though perhaps not the only or most efficient manner in every case, the provision of a singular navigational front end (e.g. taxonomy) and omnipresent search tool that collectively aggregate disparate content resources, can, from an end-user perspective, deliver the simple single point of access that many users strive for.

But therein lies the challenge to the business and IT communities. While the business side must determine the right levels of functionality needed, IT must develop approaches that simplify the delivery of such functionality and minimize the number of front-ends [5]. Understanding myriad sources of unstructured content (e.g. web content, e-mail and online files), requires the orchestration and coordination of multiple disciplines and technologies working in concert. This is the result of well-planned information architecture [5].

Pivotal points in any such information architecture are the selection and implementation of taxonomy tools, categorization tools, and search tools, each ranked as critical to information architecture by a plurality of research survey respondents.

Delphi Group’s research on user experiences with using content reveals that lack of organization of information is in fact the number one problem in information management and retrieval [4], in the opinion of business professionals. These professionals include customer service representatives, the sales team, the financial services professionals, the R&D engineers and senior executives. If these professionals are spending 20% of their time or more looking for information, then this results in an opportunity cost and represents a runaway expense item in many organizations [4].

One of the defining challenges of this era of enterprise computing is just this: How do we find the relevant and pertinent information to do our jobs and make informed business decisions? The answer is at once obvious and elusive. We must harness the computer to help manage and retrieve content at the same rate at which it allows us to create and distribute that content.

In this paper, we begin by giving an overview of taxonomy and classification technology. We then propose a framework and a system implementation for automatic e-Business Web content classification in search results. Finally, we show the results from the experiments, followed by a summary.

2. Taxonomy and Classification Technology

In this section, we take a detailed look at the state of the technology which has been developed to facilitate the creation and maintenance of taxonomies and the classification of e-Business content.

When investigating the applicability of technology to the taxonomy and classification process, it is helpful to break the process down into four distinct stages [6]. The four distinct stages of taxonomy business practice each represent a level of functionality that must be incorporated in a solution definition, and thus can potentially
be addressed via a technology implementation. These four stages are:

1. Developing of the taxonomy structure
2. Categorizing the content and placing the pointers to the documents in the hierarchical structure
3. Presenting the information (or building the interface that helps users find the information)
4. Incorporating and analyzing new content and maintaining the taxonomy structure

Taxonomy software has been developed over the past several years to increase the speed and efficiency of each of these stages. It is important to note the distinction between taxonomy design/construction and the classification of content sources and objects into that taxonomy. These are two separate and distinct operations and one should not assume that these processes are happening simultaneously or concurrently. These are serial steps. While these functions are closely related, technology solutions do not always provide both levels of functionality. More often, technology solutions focus more on the classification process [7].

There are many approaches to tackling the problem of building automatic or semi-automatic taxonomies and automating the classification process. Technology vendors’ solutions provide a variety of approaches which combine these functions in distinct configurations.

These underlying technologies are based on a number of different development approaches and algorithms, the most common being:

- Rules-based
- Bayesian
- Linguistic and Semantic
- Support Vector Machine
- Pattern Matching and Other Statistical Algorithms
- Neural Networks

2.1 Rules-based

The rules-based approach [8] is perhaps the most straightforward and user controllable approach. The rules-based approach requires experts to create and maintain a set of rules for a document to be included in any given category of taxonomy. Thus, the rules-based approach focuses more on the classification process than the construction and definition of taxonomy.

Experts define “If-Then” rules that can support complex operations and decision trees. Rule-based systems can precisely define the criteria by which a document is classified. The rule measures how well a given document meets the criteria for membership in a particular topic.

For example, a rule could be that all documents that include the terms “San Francisco,” “Chicago,” “New Orleans” be listed in a category called “Cities, USA.” Such rules often break down when ambiguous values like “Cambridge” arise. Is this Cambridge in Massachusetts or in England? Rules must be carefully articulated and made as unambiguous as possible.

Besides the content of documents, rules can be applied to metadata and even business policies. For instance, a rule might specify that only PDF documents created since January 2000 should be included in a particular category. Thus rules are a powerful and flexible means for automatically classifying content based on
not only content itself but also the metadata that describes the content’s business context (for example: author, date, or keyword data can be used in rules). The downside of rule-based systems is that expensive human domain experts have to write and maintain the rules.

2.2 Statistical Analysis

This approach supports both the creation of taxonomy and the subsequent classification of content into that taxonomy [9]. The approach measures word frequency, placement and grouping, as well as the distance between words in a document.

Typically, the statistical approach to taxonomy definition and constructions (as opposed to subsequent classification) requires some form of preliminary training. This could take the form of a basic taxonomy, defined by a human expert. But the breadth and validity of the structure, and subsequent classification rules can be automated through application of a training set of documents into the design process. In this approach, subsets of documents are identified manually and presented to the software as “exemplary” to a given topic or node of the taxonomy.

The provided sample content is analyzed and from this the taxonomy is further refined and the rules of classification established. These rules are then used to automate the analysis of new documents and their classification into the taxonomy. This approach is also referred to as “machine learning.”

Limitations of the example-based taxonomy method include problems that arise from the fact that the resulting classification is totally dependent on the breadth and precision of the training set, and the training set still must be identified manually. In any case, the statistical analysis that is performed can deploy one or more approaches: Bayesian Probability, Neural Networks and Support Vector Machines.

2.3 Bayesian Probability

The Bayesian approach [10] attempts a concept-based analysis by learning the probabilities of words being related in a given category. The Bayesian algorithm sorts documents by examining the electronic patterns contained in the text or content contained therein. Bayesian probability uses statistical models from words in training sets, and uses pattern analysis to assign the probability of correlation. This is one of the more common methods applied to building categories and taxonomy structures.

An example of Bayesian probability would be that if a given document contains the words “apples” and “oranges” it is more than likely this document is about fruit, which leads to the assumption that other fruit nouns such as “grapes” or “tangerines” will occur.

2.4 Neural Network

Neural Networks [10] create a matrix of computational nodes. These nodes track and compare topic similarity. A neural network utilizes artificial intelligence to build an interconnected system of processing elements, each with a limited number of inputs and outputs. Rather than being programmed, these systems learn to recognize patterns. Neural networks are an information processing technique based on the way biological nervous systems, such as the brain, process information. Composed of a large number of highly interconnected processing elements, a neural network system uses the technique of learning by example to resolve problems. The neural network is configured for a specific application, such as data classification or pattern recognition, through a learning process called “training.”
2.5 Support Vector Machine

Support Vector Machine (SVM) [11] is a refinement of taxonomy-by-example (i.e. it requires a training set). These algorithms are derived from statistical learning theory. SVM's calculate the maximum “separation,” in multiple dimensions of one document from another. Each document essentially a collection of words and phrases that together have meaning are represented as a vector. The direction of the vector is determined by the words (dimension) it spans. The magnitude of the vector is determined by how many times each word occurs in the document (distance traveled in each dimension). As this iterative method continuously analyses documents, it separates them into either the “relevant” space or the “irrelevant” space. By repeating the process it categorizes those documents that are “relevant” into like categories, but more importantly learns how they are different from other categories.

2.6 Semantic and Linguistic Clustering

Semantic analysis and clustering supports [12] both the creation of a taxonomy and the categorizing of content. This approach is typically language dependent. Documents are clustered or grouped depending on meaning of words using thesauri, custom dictionaries (e.g. a dictionary of abbreviations), parts-of-speech analyzers, rule-based and probabilistic grammar, recognition of idioms, verb chain recognition, and noun phrase identifiers (e.g. “business unit manager”).

Linguistic software also analyzes the structure of sentences, identifying the subject, verb, and objects. The sentence structure analysis is applied to extract the meaning. Stemming (reducing a word to its root) also helps linguistic or semantic clustering. Clustering is a technique for partitioning documents/words into subsets of similar documents/words based on the identification of common elements between the documents/words.

2.7 Combining Methodology

Of course, no single taxonomy and classification methodology is superior to another for every possible application. The trend by taxonomy software companies is to combine multiple methods to categorize the corpus of documents to increase accuracy and the relevancy of groupings. Each approach and combination of approaches has pros and cons associated with them, and their use depends on a design perspective and performance characteristics desired. The bottom line is to understand how these differences affect system performance in the only environment that matters user’s unique data environment.

3. Framework for Automatic e-Business Web Content Classification

As the above outlines, various issues are related to classification technology. One of the major objectives of classification technology is to classify any type of e-Business Web content. This is a difficult task since it required a precise search over the internet. The speed of the classification is limited by the skills and number of individuals assigned to this classification task. Addressing these issues of taxonomy definition, construction and classification must have a dual focus:

- Identify user’s e-Business content requirements.
- Understand and evaluate available technology alternatives.

3.1 e-Business Web Document Collection

For the purpose of information classification in the e-Business Web
documents and with a view to facilitate the implementation of the classification system, a more formal definition of e-Business is followed. It helps to decide clearly which “electronic business activities” fall into a well-defined, quantifiable framework of e-Business. Thus, an implicit distinction of e-Business from the structures of the traditional commerce is required, since new ways are invented in order to measure the revolutionary elements and methods. In an attempt to investigate and measure the ways e-Business has fundamentally changed the way of transactions, a sharp criterion in distinguishing the e-Business pages is taken under consideration.

3.2 Determining Taxonomy and Classification Technology

It is therefore critical to precede any investigation of technology solutions for taxonomy and classification with a determination of what user’s needs are, the types of content that will be managed, and the best approach to exposing the taxonomy to the user.

It is important to understand the approaches available to taxonomy and classification presentation and integration. Technology tools are available as stand-alone applications and as components to integrated knowledge management and document management systems.

When investing the technology alternatives pay attention to the user interface. Does the tool come with a predefined interface? Is this interface conducive to the way users think about their content? Nested file folders that mimic the popular MS Windows GUI are available, as are nested tree structures, alphabetical listings of topics and sub-topics, tab based interfaces, heat maps, hyperbolic trees that resemble tinker-toy like connections of topics, and even voice recognition interfaces. A number of vendors view this technology as a fundamental component of the information infrastructure. Just as relational databases are a fundamental infrastructure component of applications such as accounting, CRM, and other enterprise applications, taxonomy software can be the infrastructure component that correlates unstructured data. This design philosophy positions taxonomy software as a core module in an architecture that works on the unstructured data within the organization.

However, using a vendor-provided interface or creating one, the technology should provide some approach to integration. Minimally, it is anticipated that the taxonomy functionality will be integrated with search tools.

4. System Implementation

From the previous section, we proposed the automatic classification e-Business content framework. To implement this framework, we collected 4,250 e-Business Web documents from CMU World Wide Knowledge Base [13] as our target group. The reason why we focus on this target group is Web page mining task is a complex process and one needs to address several issues. That is the reason that most papers deal with only a specific set of web documents. The tasks that need to be addressed are information retrieval, removal of static and dynamic noise, using the structure to find the importance of the information, extracting the actual information from unrelated data and storing it. The process can be accomplished if we filter the web documents several times using different aspects and then only certain data will be left which will actually be the minimal integral data that the web site uniquely represents.
4.1 Defining e-Business Web Content

Following the concepts of the Business Media Framework (BMF) (Klose, Lechner and Ulrike, 1999) and e-Business model can be analyzed into a series of concurrent processes, while these processes are implemented by elementary transactions. According to the accepted transaction cost theory, four phases distinguish an e-Commerce transaction. The knowledge phase, the intention phase, the contract phase and the settlement phase. The distinction of these four phases serves as an analytical tool in order to classify e-Business Web content.

4.2 Defining Classification Method

In this system, we used SVM (section 2.2) as our classification method. The reason why we focus on this method is because it is simple and fast to implement. The main processing steps in this method involve:
2. Execute word segmentation.
3. Create the e-Business Vector then do normalization.
4. Find the average of all vectors for each group.

In this system, we used 1250 e-Business Web documents as a training set. The training set is classified by manual into 7 groups. The outputs from the system are average vectors that represent for each e-Business Web content group. Figure 1 presents the details in the e-Business Web content classification processes.

Figure 1 e-Business Web content classification processes
5. **Experimental Result**

In order to evaluate the system performance in e-Business Web document classification, Offline test sets (2000 e-Business Web documents) were taken under consideration. This experiment used a relevance threshold in the similarity value between test sets vector and average vector as explained in Section 4.2. For online test, we test within search results from the Google Website. We found that average system performance is about 85%.

6. **Conclusion**

In this paper, we propose a framework and a system implementation for automatic e-Business Web content classification in search results. In order to evaluate the system, two test sets (offline and online) were taken under consideration. For offline testing, we used 7 e-Business Web collection groups from CMU World Wide Knowledge Base, 1250 Web pages for training and 2000 Web pages for testing. For online testing, we used Web collection from the results in search engine for testing. Both of the results show that the average system performance is about 85%.

However, this framework can be recommended in the experiment of another algorithm to improve performance (e.g. statistic algorithm, neural network, etc.).

7. **References**


