Making Sense of Search Results by Automatic Web-page Classifications

Ben Choi
Computer Science
College of Engineering and Science
Louisiana Tech University, USA
pro@BenChoi.org

Abstract: This paper reports the development of a system for automatically organizing Internet web pages into meaningful categories. The aim of the system is to allow Internet users to find useful information in less time. The current problem with using the Internet is how to find the information that we need. With the explosive growth in the Internet, the information overload situation is getting worse. The proposed system automatically classifies web pages based on three types of information: (1) The system analyzes organizational information among web pages (inter-web-page relationship), such as an URL and links within a web page. (2) It analyzes the meta-web-page information such as data contained in META tags and formatting data of a web page. And (3), it analyzes web-page-content information such as keywords and phrases in the content of a web page. Our results show that combining all three types of information provides better accuracy.

Introduction

Although we know that "information is power," the current problem with using the Internet is how to find the information that we need. With the explosive growth in the Internet, the information overload situation is getting worse. Since currently the Internet technology is one of the most important sectors in our economies, it is of central importance for our society that we take action to address this information overload problem. Nowadays, a simple keyword search on the Internet may return thousands of web pages (even up to a million in one of our tests). To find the right information, one may have to scan the outlines of the first hundred web pages (and usually give up after that). To decrease the amount of time spent searching for the right information, it is of central importance that we take action now, before the situation becomes unmanageable.

One technique for managing vast amounts of information on the Internet is to arrange web pages into categories. From the point of view of web-page classification, search engines can be grouped into three classes: manual classification, non-classification, or automatic classification. The current status is that there is one search engine providing manual classification (www.yahoo.com), one providing automatic classification (www.northlight.com), and the rest of them either provide not classification or rudimental manual classification (AltaVista, Excite, Go, DirectHit, Google, and Lycos). Although there is already one automatic classification system available in the market, the proposed system addresses new issues (such as inter-web-pages analysis) and provides enhanced implementations (such as object-oriented data organizations and distributed processing).

The automated categorization of web documents has been investigated for many years. For example, Northern Light received a United States patent on July 13, 1999 for their classification mechanisms (Krellenstein 1999). Mladenic (1998) has investigated the automatic construction of web directories, such as Yahoo. In a similar application, Craven et al. (1998) intend to use first-order inductive learning techniques to automatically populate ontology of classes and relations of interests to users. Pazzani and Billsus (1997) apply Bayesian classifiers to the creation and revision of user profiles. WebWather (Joachims et al., 1997) performs as a learning apprentice that perceives user's actions when browsing on the Internet, and learns to rate links on the base of current page and the user's interests. For the techniques of construction of web page classifiers, several solutions have been proposed in the literature, such as Bayesian classifiers (Pazzani & Billsus, 1997), decision trees (Apte et al., 1994), adaptations of Rocchio's algorithm to text categorization (Ittner et al., 1995), and k-nearest neighbor (Masand et al., 1992). An empirical comparison of these techniques has been performed by

Choi, Ben (2001) "Making Sense of Search Results by Automatic Web-page Classifications",
Proc. of WebNet 2001 -- World Conference on the WWW and Internet, pp. 184-186.

Pazzani and Billsus (1997). The conclusion was that the Bayesian approach leads to performances at least as good as the other approaches.

Our Approaches

Our project addressed the following four major issues for automatic classification of web pages:

(1) Develop a process and a database system to allow arrangement and storage of hieratical categories: Our strategies include using Object-Oriented database and storing the hieratical categories in a tree structure.

(2) Develop an analysis system to determine whether a web page belongs to a specific category. The system analyzes the follow three types of information: (a) Inter-web-page relationships including URL (Universal Resource Locator) and links contained in a web page: the analysis includes going down the links pointing out of a web page and checking the given URL of the current web page for information such as the type of resources and web site organization (e.g. www.cnn.com/health indicates that the web pages contained in the health directory may be health related.). (b) Meta-web-page information including HTML Meta tags and formatting data: the analysis includes checking META tags such as tags for description, keywords, section, subsection, and date. In addition, the formatting data of the web page are also analyzed (e.g. table of content or resume usually formatted in certain way). And (c), web-page-content information including keywords and phrases: the analysis includes matching keywords within a web page to keywords for a specific categories. Words appeared within some HTML tags carry more weight than others.

(3) Develop a mechanism that can learn from training web pages to identify attributes of a specific category. The learned and programmer provided attributes are then used by the analysis system. Training web pages and their categories are provided to the learning mechanism. All training web pages are initially tokenized, and tokens shorter than three characters are removed. The set of tokens (words) is filtered to remove HTML tags, punctuation marks, numbers, and stop words, such as prepositions, articles and conjunctions. Moreover, a stemming algorithm is applied to remove suffixes such as -s, -es and -ies, to preventing separate frequency counting for those words that differ in the number (e.g., computers and computer). Separate counting leads to flattening frequency histograms and has the effect of making relevant and irrelevant features less distinguishable. A selection criterion is developed to select the tokens that are best representative of the categories. Normalization methods are also used to account for difference in the number of words in the training web pages. Methods to learn from inter-web-page relationships and from meta-web-page information do not yet exist, but are being developed in this project.

(4) Develop a strategy for the proposed system to be able to port to parallel and distributed hardware platforms: Our strategies include using CORBA (Common Object Request Broker Architecture) interfaces. CORBA specifies the functions and interfaces of an object Request Broker (ORB), which act as an object bus that allows remote and distributed objects to interact.

Automatically organizing Internet web pages into meaningful categories is only a small step toward addressing the information overload problem due to explosive growth in the Internet. Other promising strategy includes using intelligent agents that learn user's behaviors and preferences. Based on the profile of the user, the intelligent agents then automatically gather information for the user.

References

Apte, C., F. Damerau, & S. M. Weiss (1994). Automated learning of decision rules for text categorization. *ACM Tans. on Information Systems*, 12(3), pp.233-251.

Craven, M., S. Slattert, & K. Nigam (1998). First-order learning for web mining. *Lecture Notes in Artificial Intelligence*, 1398, pp.250-255, Springer: Berlin.

Ittner, D., D. Lewis, & D. Ahn (1995). Text categorization of low quality images. *Symposium on Document Analysis and Information Retrieval*, pp.301-515.

Joachims, T. (1997). A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization, *Proc. of the 14th International Conference on Machine Learning*, pp.143-151.

Krellenstein, Marc F. (1999). Method and apparatus for searching a database of records, US Patent & Trademark Office, United States Patent 5,924,090, July 13, 1999.

Massand, B., G. Linoff, & D. Waltz (1992). Classifying new stories using memory based reasoning. *Proceedings SIGIR'92*, pp.59-65.

Mladenic, D. (1998). Turning Yahoo into an automatic web-page classifier. In H. Prade (Ed.), *Proc. 13th European Conference on Artificial Intelligence*, pp.471-474.

Pazzani, M. & D. Billsus (1997). Learning and revising user profiles: The identification of interesting web sites. *Machine Learning Journal*, 23, pp.313-331.