Exploiting Machine Intelligence

Precise Information Cataloging Solutions™

GammaSite White Paper
May 2001

Table of Contents

Introduction

In the increasingly competitive global business environment, organizations are looking for ways to decrease costs, increase productivity and achieve greater scalability.

Organizations are now facing two new challenges:

1. *Tasks have become more knowledge intensive* – More of the basic functions within a company are becoming dependent on the employees' ability to easily find and access relevant information. These "knowledge workers" are spread throughout the length and breadth of a company from sales and marketing to research and development and depend on finding relevant information in order to make decisions. The proliferation of knowledge workers who are dependent on a constant flow of information has forced businesses to invest heavily in information systems.

2. *The information glut* – A combination of communication and production technologies has created a mountain of information. Companies are bombarded by a constant barrage of information from external sources such as the Internet, news wires and publications. In addition, companies themselves are producing thousands of internal documents on a daily basis. As a result, the relevant information needed for decision-making is "masked" under this mass of information. The information glut has led many to the following conclusions:

    1) Human beings cannot manually manage the vast amount of information in an organization

    2) Companies today lack the tools to effectively manage and organize the information within their company

Poor information management and retrieval impact a company's bottom line. Strategic decisions based on incomplete information and are all symptoms of a company that is failing to efficiently manage and leverage its information resources.

With the help of Information Cataloging solutions, companies leverage the power of information within an organization and turn information masses into valuable assets.

GammaSite employs state-of-the-art machine intelligence concepts to automatically categorize information into a pre-defined and detailed taxonomy structure. A taxonomy, or topic tree is a hierarchical presentation of information that represents a specific knowledge area. On top of such a taxonomy structure, GammaSite provides search capabilities, effective user and community profiling routines, and user-friendly tools for defining, updating, refining and filtering the taxonomy structure and content.

The Importance of Cataloging Information

One of the key business challenges facing organizations is the ability to transform growing information masses into a competitive advantage. One of the ways in which this can be turned into reality is to organize this growing amount of information into a taxonomy in which it is easy for end users to access and retrieve relevant information.

Some benefits for cataloging information within an organization include:

1. Improving productivity by providing knowledge workers with the ability to **easily access and retrieve relevant information** according to topic, thus significantly reducing time spent searching for information

2. **Enable better informed and quicker decision-making** by having all available information on hand instead of within multiple areas within the organization

3. **Leverage internal knowledge assets** of the company.

Organizing information within a taxonomy structure enables end users to intuitively browse through a listing a relevant subtopics within the taxonomy, rather than just using keyword search. An inherent problem of key-word search is that users may experience difficulties in defining appropriate queries and therefore spend excessive amounts of time trying to formulate well-posed queries and are thus forced to sift through a large list of irrelevant answers.

Types of Information Cataloging Approaches

While there is a consensus that cataloging information is a necessity within organizations, there are varying methods being employed currently in the marketplace. GammaSite describes below the most common ways to organize information including manual classification, rule-based, linguistic and machine learning with emphasis on the advantages of GammaSite's machine learning classification system.

Manual Classification

In this method, subject experts manually review documents and then organize them according to the topics within the existing hierarchical topical structure. While this method is appealing due to its perceived accuracy, it has several shortcomings including significant amount of labor needed to categorize and maintain these information directories as well as inconsistencies in classification results among multiple people classifying documents. This type of solution is not only very costly, but is also not scalable and cannot cope with the immense amount of information that flows into and within a company.

Rule-Based

This approach is based on a pre-defined set of rules that characterize the different concept classes of interest. The rule-based classifier can be based on simple rules (e.g., 'does a certain key-word appear in the document more than a pre-specified number of times'), or sophisticated operators (e.g., 'does a certain word appear in italics near a section title'). The advantage of rule-bases approaches is the possibility of incorporating prior knowledge into the classifier. For example, suppose we are constructing a classifier for patents related to cellular phones. Besides the rules that identify the topic, we might also use our knowledge of the domain and require that the document contain a section titled "Claims". The main drawbacks of rule-based classifiers are their extreme reliance on man-made

rules. The specification of such rules is typically performed by knowledge engineers whose expertise is used to quickly and effectively learn the field of interest (as well as related fields) and extract the relevant rules. This process is extremely costly in terms of man-hours.

The resulting classifier is sometimes too coarse to capture subtle topic differences, resulting in low accuracy. Moreover, the constructed classifiers are highly biased and depend strongly on the particular person's prior knowledge and talent. In this sense, the construction of rule-based classifiers can be thought of as more an art than a science.

A good lesson regarding rule-based classifiers can be learned from the related field of handwriting recognition. Initially, during the early 1960's, many attempts were made to use knowledge engineering in order to form rule-based classifiers for handwriting. These attempts failed to achieve sufficient accuracy. The state-of-the-art in handwriting classification today is based on statistically learned classifiers.

Linguistics

Several attempts have been made to incorporate linguistic knowledge into text categorization systems. Such Natural Language Processing (NLP) approaches attempt to make use of morphology, syntax and semantics in order to construct efficient data representations from the raw text, and also to resolve semantic ambiguities that inherently exist in any text. While these techniques are successful in a number of more constrained problems such as 'word sense disambiguation', the current consensus in the field of textual pattern recognition is that these techniques have not yet matured to the level where they can compete with the state-of-the-art statistical classifiers. In fact, among the more successful NLP algorithms are those based on statistics. For instance, the problem of constructing a 'semantic map' of words and their synonym and hypernym connections (such as WordNet) can be automatically achieved using unsupervised statistical methods, while a manual construction requires a monumental amount of labor.

Statistical Machine Learning (SML)

The statistical approach to classifier construction is based on the assumption of an underlying statistical regularity. This assumption is always valid in practice when the data emerges from a "natural" source. Within this framework, one builds a set of alternative statistical models and attempts to find the best model, or a weighted combination of models, based on a finite set of examples. A particular advantage of the statistical approach to classifier construction is the availability of a large body of work within the field of Statistics, Information Theory and Computational Learning Theory, which provides a principled and systematic approach to the problem of model selection and validation. It should be emphasized that while the approach is data-driven, the incorporation of prior knowledge through model construction provides users with sufficient flexibility to incorporate domain specific knowledge.

The main advantages of the SML framework are its generality and flexibility. This means that SML can be (and in fact has been) successfully applied across a broad spectrum of problems. An appealing feature of SML-based classifiers is their guaranteed generalization ability from observed data ("training data") to novel data, given the plausibility of the underlying statistical assumptions (which can always be modified if the need arises). No such guarantees are available for the rule-based and semantic/linguistic approaches. Using SML it is possible to construct accurate and robust classifiers based on relatively small training sets. The labor costs involved in using SML involve constructing the initial statistical models and deciding on the best feature representation and on the gathering of appropriate training sets.

GammaSite's Machine Learning Approach

GammaSite uses the most advanced learning algorithms and classifiers in order to learn an appropriate classifier for each category in the taxonomy. These learned classifiers are based on automatically generated features. The resulting classifiers are tolerant to noise and lead to very robust classification in comparison with Boolean query type classifiers, which are known to be highly susceptible to noise and outliers. Moreover, it is a simple matter to modify the classifiers based either on new information that becomes available or on new examples that may help to sharpen their performance and accommodate organizations' changing needs. The features used by the classifier at each node are both semantically meaningful and context sensitive, while retaining robustness to noise. This enables Gammasite to retain the advantages of statistically robust and semantically meaningful classification, a feature that is very difficult for other classifiers to maintain.

Advantages of GammaSite's Machine Learning Approach

1. The emphasis in Gammasite's approach is on constructing solutions to a complex task by learning to generalize from a finite set of examples, rather than forming an algorithmic solution based on some prior knowledge and a set of heuristics. This approach has been shown to provide much more robust and flexible results, in comparison to the brittle solutions often devised within the classic Artificial Intelligence paradigm.

2. The main assumption underlying GammaSite's approach is statistical in nature, making use of advanced tools from the highly developed field of Statistics, which allows for a rigorous and effective characterization of the results. The learning approaches taken permit the effective utilization of prior knowledge and its incorporation into the learning process. However, only very rudimentary prior knowledge needs to be used as the system can learn to generalize based on the data.

3. Utilization of statistically meaningful and contextually relevant features, based on state-of-the-art unsupervised learning approaches.

GammaSite's Solution

GammaSite has developed a comprehensive solution for the precise cataloging of text and hypertext. The entire creation process is based on sophisticated algorithmic ideas and on relatively complicated software modules. See below for an overview of the main components of GammaSite's solution.

General Overview

Figure 1 depicts a high level view of GammaSite's system. The automatic categorization process begins by constructing a subject specific taxonomy and inserting a few example documents for each desired sub-topic. This is performed using GammaSite's taxonomy builder, Expert Tool. This step is the only element in GammaSite's system which requires manual labor. This manual step enables the utilization of a small amount of human expertise and allows the computer to achieve results that are close in accuracy to human categorization performance. The information provided during this stage allows for the automatic construction of both a focused crawler and a categorization machine that includes a large set of expert classifiers. After its creation, the focused crawler constantly crawls internal and external data sources for information relevant to the taxonomy. The automatic categorization module then classifies any given document into its correct category(ies) within the taxonomy. Each new document is then passed through an automatic summarizer, which generates a concise abstract of the document and is then integrated into the search engine to enable powerful search capabilities.

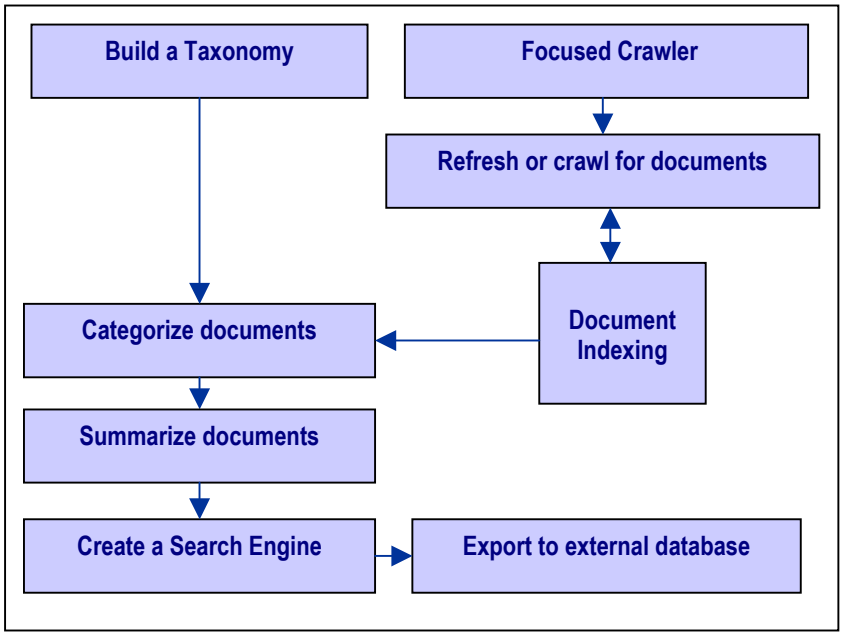


Figure 1: Overview of GammaSite System

Constructing a Taxonomy

There are several steps involved in constructing a taxonomy. After the subject is defined, and the appropriate sub-topics are specified, each category within the taxonomy needs to be given a few examples of similar types of documents that the user seeks to catalog. Although it is advised that an expert in the relevant field should perform the taxonomy construction, GammaSite provides a proprietary taxonomy builder, Expert Tool, which provides suggestions, statistical parameters and consistency checks during the taxonomy construction process.

While some attempts have been made at automatically constructing a taxonomy using self-organization of a large pool of unstructured data, these attempts typically achieve unsatisfactory results. The complexity of the problem and the detailed expert knowledge required for a non-trivial topic precludes a fully automatic solution of this task.

Data Representation

A problem of major importance that needs to be addressed in dealing with complex data formats (e.g., text, images, video etc.) concerns data representation. For example, a common approach in the field of text categorization is to represent a document as a high-dimensional vector, where each component of the vector represents (say) the frequency of the appearance of a certain word in the document. This simplistic approach, called the "bag-of-words representation", can be effective in some domains, but suffers from an inherent statistical sparseness problem in natural documents. At GammaSite, we have developed a sophisticated data representation procedure based on the most recent research in the field of unsupervised learning and contextual clustering. It is important to stress that the representation is node-dependent. In other words, the same document appearing at two different nodes may possess different representations. For example, consider a node in a hypothetical tree dealing with transportation. A document about cars may contain the feature term *engine* when attempting to separate the sub-class of *motor transportation* from the class of *animal transportation*, but is irrelevant in separating the class of *four-wheel vehicles* from the class of *motorcycles*.

Training, Classifier Design and Categorization

One of the biggest challenges in the fields of pattern recognition and statistical machine learning is the problem of designing classifiers with low "sample complexity". Theoretical bounds based on VC-theory [Vap98, BMM98] and empirical results presented in professional conferences and journals report on prohibitively large "training sample sizes" required for achieving acceptable classifier accuracies.

GammaSite has developed methods for constructing a sophisticated system of classifiers that achieve high accuracy based on *impressively small training sets*. This enables minimization of the manual work required to construct a taxonomy, thereby providing substantial savings in time and costs.

The construction of our classifier system is based on several ideas that exploit the hierarchical organization of data. For instance, the construction of an effective classifier at each node in a hierarchical taxonomy is greatly facilitated by the high-quality and flexible representation constructed by GammaSite's hierarchical data representation algorithm. This algorithm can form a compact representation that permits the utilization of the most advanced classification algorithms. GammaSite has developed a class of sophisticated Bayesian Kernel Classifiers (***BKC***), leading to robust and flexible algorithms. Moreover, GammaSite has developed methods that deal with particularly noisy data, to a large extent eliminating the problem of over-fitting that plague many other learning algorithms. Finally, sophisticated mathematical ideas were developed for the purpose of categorizing a given object into the taxonomy [EFT97]. Due to the sheer size and complexity of taxonomies for organizations (some taxonomies may have thousands of categories and more than eight levels), GammaSite developed efficient categorization procedures, which are able to access a document and classify it very quickly and accurately. In addition, since real-world data is inherently multi-labeled, namely each document may naturally belong to several sub-topics simultaneously, GammaSite has developed patent-pending procedures that are able to find the most appropriate labels for each input document.

Focused Crawling

Efficiently collecting topical documents is of paramount importance in the context of large enterprises attempting to systematically and reliably expand their databases. The hardware facilities required for effective general-purpose crawling are immense. GammaSite has developed a patent pending focused crawler that allows for the most efficient focused crawling within any given topic domain. Our focused crawler traverses the net while avoiding unrelated links. This type of crawling allows for quick and impressive coverage of the Internet within each topic domain. The basic idea is to learn *online* which links lead to the most successful documents, thereby eliminating the number of attempts to move towards irrelevant sources of information. The graph in Figure 2 *qualitatively* demonstrates the effectiveness of this type of crawling. While focused traversal can gather close to 100% of the relevant information within a short traversal, the entire web must be traversed in order to achieve the same coverage using a traditional non-focused traversal.

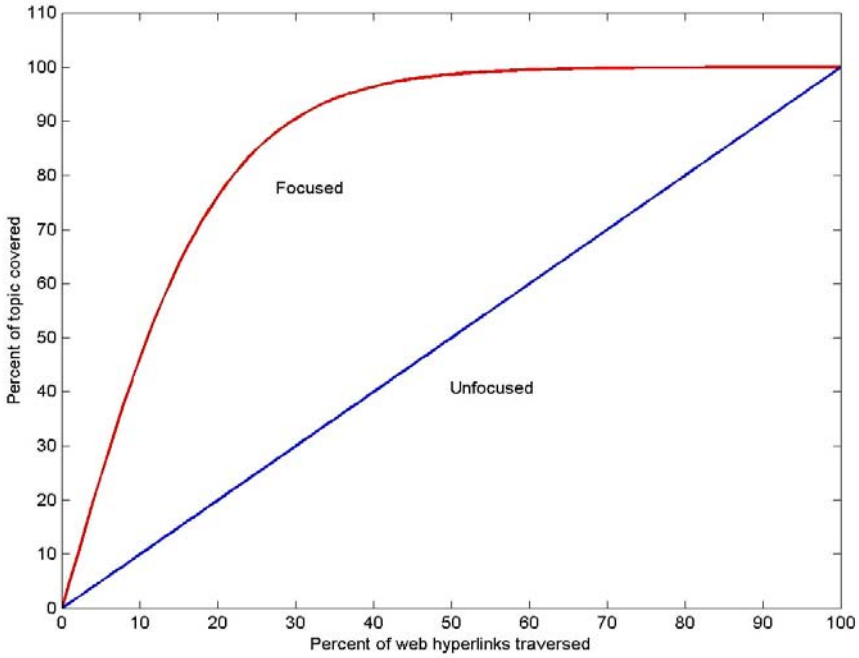


Figure 1: Focused versus Unfocused Crawling

Summarization

Documents within a taxonomy are often of a complex form, either in terms of length or in terms of data formats. Thus, in presenting a list of documents for the user's inspection, it is beneficial to attach a short summary of the page. This feature is important in helping the user decide on the relevance of a given document to his needs without reading the whole document. GammaSite employs summarization algorithms that are able to extract from a given document a desired number of sentences that most effectively summarize the content of the page. The algorithm is based on parsing the page into its basic components, and making use of various textual cues such as the position of a word in the page, the font used, the relationship to a set of keywords provided by the information expert at each node, as well as many other cues.

Case Study – Web of Golf

Overview

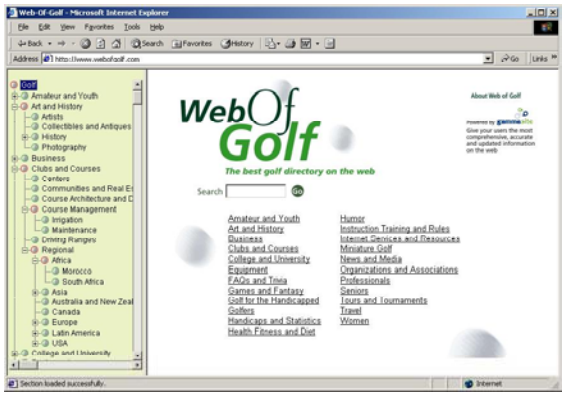
To demonstrate GammaSite's precise automatic categorization solution, GammaSite created a taxonomy on the topic of Golf, called the "Web of Golf".

How "Web of Golf" was Built

Once the Golf taxonomy was manually defined, the sub-topics identified, and a minimum number of sample documents given to the system, GammaSite's topic-specific focused web crawler was activated and retrieved over 500,000 **relevant** pages. Within a few hours, these relevant pages were automatically categorized within a detailed taxonomy structure of over 300 subtopics while spam and duplicates were automatically filtered out.

Unlike traditional human edited taxonomies, the cataloging of documents for the Web of Golf is automated, enabling comprehensive coverage of the entire Internet while preventing "link rot". The Web of Golf automatically categorizes pages by subject and is topic specific, ensuring that users receive highly relevant and focused results in response to queries.

The result of this work is the Web of Golf, the most complete and precise golf directory on the web.



To view the Web of Golf, go to www.webofgolf.com.

References

[But00] Butler, D., Souped-up search engines, Nature, Vol. 45, pp. 112-115, 2000

[BMM98] Bartlett, P., Maiorov, V. and Meir*, R.,* Almost linear VC dimension bounds for piecewise polynomial networks, *Neural Computation*, 10(8): 2159-2173, 1998

[BP98] Brin, S. and Page, L., The anatomy of a large-scale hypertextual Web search engine, In *Proceedings of the 7th International World Wide Web Conference*, pp. 107-117, 1998.

[CT91] Cover, T. and Thomas, J. *Elements of Information Theory*, John Wiley & Sons, 1991

[DE00] Bulletin of the Technical Committee on Data Engineering, Vol. 23(3), pp. 1-48, *Special issue on Next-Generation Web Search, IEEE Computer Society*, September 2000

[EFT97] El-Yaniv, R., Fine, S., and Tishby, N. Agnostic clustering of Markovian sequences, In *Advances in Neural Information Processing Systems*, Vol 10, pp 465-71, 1997.

[Gue00] Guernsey, L., The search engine as cyborg, New York Times, June 29, 2000

[Kle99] Kleinberg, J., Authoritative sources in a hyperlinked environment, In *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*

[MEB00] Meir, R., El-Yaniv, R. and Ben-David, S., Localized Boosting, In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory, 2000.*

[MS99] Manning, C.D. and Schutze, H. *Foundations of Statistical Natural Language Processing*, MIT Press, Boston, 1999.

[Nea96] Neal, R. *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics No. 118, Springer-Verlag, 1996.

[She00] Sherman, C., The future revisited: what's new with web search?, http://www.onlineinc.com/onlinemag/OL2000/sherman5.html, May 2000

[Vap98] Vapnik, V., *Statistical Learning Theory*, John Wiley and Sons, 1998.

GammaSite, Inc.

9 Maskit
Hertzliya, Israel 46733
Phone: 877-678-5049
Fax: 972-9956-9958

http://www.gammasite.com

For Sales and Product Information
info@gammasite.com

For Partnerships
partners@gammasite.com