Automatic Tagging and Categorisation:

Improving knowledge management and retrieval

1. Introduction

Unlike past business practices, the modern enterprise is increasingly reliant on the efficient processing of unstructured information. It is estimated that this unstructured data doubles every few months (Gartner). By automating knowledge tagging and categorisation, iLevel Software enables the automation of business tasks that were previously handled manually, improving the discovery and retrieval of valuable information independently of where it resides. This document describes iLevel Software's approach to automatic tagging and categorisation, and shows a comparison with other alternative methods.

2. The iLevel Software solution

iLevel Software develops and markets automatic categorisation and tagging solutions based on proven software, used by large enterprises, publishers and governments. The company is leading the drive for intelligent, automatic information management that adds meaning and context to unstructured text. The tagging and categorisation engine automatically classifies documents into categories defined in a taxonomy, using Statistical Machine Learning. Documents are automatically tagged with descriptive keywords and category paths, enhancing searchability; information repositories, file servers or remote web site pages are browsed using a tree of links organised into related topics; content is distributed to staff, partners or customers according to their interest in specific categories.

The tagging and categorisation software easily integrates with your enterprise search engine, corporate portal, content or document management system, CRM, knowledge management and data mining applications, enhancing their usefulness for employees.

a. Scientific foundations

The iLevel Software categorisation solution is powered by the GammaWare tagger, which uses Statistical Vector-Supported Machine Learning to intelligently associate text with categories. Based on Gammasite's patent-pending algorithms, the software is able to achieve the highest level of accuracy with a very small set of examples per category, when compared to other products in its class.

b. The process

i. **Taxonomy creation**: (a tree of categories residing under one central topic). Taxonomies can be created or imported, and then edited using the Taxonomy Manager. The Category Suggesting Tool can assist in taxonomy building, by suggesting a structure for large pools of documents in a repository.

ii. **Preparing training sets**: A training set is a group of documents (approx. 5), which represent strong positive examples for a single category. Negative examples for the category are taken from the training sets of its neighbors in the tree.

iii. **Training the software**: the engine analyses training documents using Statistical Vector-Supported Machine Learning, and automatically creates a classifier for each category. A classifier is a software function that can analyze a document, determine whether it belongs to a specific category, and return a confidence score (ranking). The higher the score, the higher the software's confidence that the document belongs in the category.

These initial steps need only be performed once per taxonomy. Once training sets have been assembled and classifiers trained, any document can be categorized into the taxonomy with no human intervention.

iv. **Fine tuning**: the engine allows you to perform "dry runs" on the training documents, and provides detailed information that helps to improve the accuracy of the classifiers. Fine tuning can take the form of adding, removing or editing training documents, adding or editing discriminating words and phrases, or setting category-specific parameters.

v. **Categorisation**: Documents are fed into the engine, one by one, or in batches. Each document is analyzed by classifiers of the taxonomy's first-level categories. If the document receives a high enough confidence score, the software attempts a finer classification, deeper within the taxonomy. At the end of this process, the software returns one or more category matches for each document, and a human operator can opt to view, edit and approve the category paths. Based on these category matches, the engine can also generate keywords and summaries for the document.

c. Features and benefits

i. **Automation:** given the uninterrupted growth of unstructured information, efficient processes that manage and extract value from information are only dependent on the ability to automate the tasks that previously have been performed with manual labour. iLevel Software's tagging and categorisation engine eliminates previously expensive and innacurate manual labour, by automating a wide number of business tasks, and in doing so, delivering significant bottom-line savings.

ii. **More accuracy:** iLevel Software uses the GammaWare engine for its underlying categorisation capability, which currently offers the highest precision and recall rates in the automatic categorisation market. Precision is the percentage of documents placed in a category that actually belong to that category. Recall (or "coverage") is the percentage of documents in the entire repository that belong to a given category, and that were classified correctly by the software. The tagging engine is not only best at getting the right classification, but it also misses fewer documents when building a category. In benchmark tests checking both precision and recall, the tagging engine attained significantly higher scores than competing products. These quality results are made possible by GammaSite's state of the art, patent-pending machine-learning algorithms.

iii. **Crawls existing information repositories, such as file servers and websites**: the engine will allow indexing of local file servers as well as remote websites. Content is classified and can be retrieved or browsed by topic, irrespective of its current residing location.

iv. **Ranking/relevance**: concepts found in documents determine whether each should be associated with one or more categories. Within each category a document can have a different relevance, therefore the associated ranking is used to influence how high the document appears in retrieval lists.

v. **Keyword and summary generation**: the engine uses the results of automatic categorisation as a basis for generating context-sensitive keywords and summaries. In many applications, keyword and summary generation is extremely important, because it allows for automatic tagging of documents with meaningful metadata.

   A context-sensitive summary is a portion of a document that describes it briefly in context with the category it was classified into. The categorisation engine often generates more than one summary for a single document: one for each category it was matched to.

Context-sensitive keywords are words or phrases that describe a document in context with its subject. For example, a document about Nelson Mandela, categorized under “Society / Human Rights / South Africa,” could be aptly characterized by keywords like “apartheid,” “Robben Island prison” and “truth commission.” Keywords about Mandela’s family, friends and hobbies would be less useful, in the context of the category.

vi. **Faster to implement and use**: due to its sophisticated machine-learning approach, the engine can achieve very accurate categorisation with minimum setup. It requires around five example documents per category, compared to twenty or more for other solutions in its class. This makes the tagging and categorisation engine significantly less labour-intensive than competing products. Furthermore, unique features of the software such as the Category Suggestion Tool reduce the time needed to manage categorisation on a daily basis.

vii. **Language-independence**: it is very important for organisations nowadays to be able to deliver the right information to the right audiences, independently of the language the content is represented in. The tagging and categorisation engine is language-independent. It does not rely on any intimate knowledge of english grammatical structure, of in fact of any other particular language. It treats words as abstract symbols of meaning, deriving its understanding through the context of their occurrence rather than a rigid definition of grammar.

viii. **Exceptional speed and scalability**: iLevel Software’s solutions are deployed to solve mission critical business problems. Rigorous requirements often imposed on the technology demand that the software provides high-performance, high capacity and a scalable platform for content exploration.

ix. **Security**: the tagging engine doesn’t store any content – it simply receives XML-encoded requests from a content crawler (that scans websites, file servers or database content), and returns categorisation results. This means the system poses no security risk, since you can continue using existing security configurations and the indexed documents stay where they are. In contrast, many categorisation solutions store documents, protecting them with a proprietary security scheme. Those approaches provide questionable protective measures for more sensitive material, while making integration more complex.

x. **Solves the hierarchical recall problem**: GammaWare offers the only solution that circumvents the Hierarchical Recall Problem, which can dramatically reduce the accuracy of automatic categorisation. The problem stems from the fact that in general, categorisation software usually attempt to filter documents down the taxonomy tree, matching them to categories one level at a time. GammaWare solves the hierarchical recall problem by predicting the highest probability for a document to belong to a relevant sub-category, using specially-developed statistical algorithms. Documents belonging to categories deep within the tree are classified without filtering down the levels. The result is dramatically improved recall, which makes for better categorisation results.

d. Information browsing and discovery:

i. **Taxonomy/topic-based navigation**: corporate data repositories, intranets and websites are enhanced by automating the creation and maintenance of topic-based Yahoo!-style directories of information. With the iLevel Software solution, users drill down these web interfaces in order to find exactly the information they need, as well as related data.

ii. **Search and retrieval** - **metadata and full text**: complementary to automatic tagging of documents and to enable quick and accurante retrieval, all content and metadata are indexed for fast searching. Search capabilities include Boolean and proximity operators, category or metadata field-based searches, natural language searches, wildcards, vector space queries, ranking, XML schema-based searches, multilingual searches, and more.

e. **Integration**: the tagging engine offers a powerful and well-documented API, which allows its easy integration with your existing enterprise search engine, corporate portal, content/document management system, CRM, knowledge management or data mining applications. These can become "clients" of the tagging engine, transmitting documents to the server and receiving the respective categorisation results. All communications take place in eXtensible Markup Language (XML).

3. Alternative approaches

Many companies claim to have solutions that solve the challenge of managing unstructured information. However, most of these systems and approaches have severe limitations particularly where scalability and costs are concerned. For example:

a. **Keyword searches of boolean query**: the most traditional approach to information retrieval is through traditional keyword search. This simple method involves users typing keywords into a text field. A search engine scans a list of documents and returns a list of those containing the search terms. This method offers a lot of irrelevant results, no context and is highly innacurate, as the same words can have different meanings depending on the subject (e.g. the word “chip”). It also requires intensive user participation and manual intervention, which in environments with large amounts of information can be extremely time consuming, error-prone, and therefore costly.

b. **Manual tagging**: creating taxonomies that address various information types (including documents, structured data, html, media and XML) is imperative. Manual tagging schemes are becoming an increasingly popular method of labeling digital material. However there are significant barriers to ensuring the increase in efficiency of managing information, and therefore keeping costs low. Very often after some time of using a manual tagging method, all documents end up being categorised as “General”, which reduces the accuracy of retrieval. Inconsistency is also a drawback of manual tagging since humans tend to describe information based on their own experience, and on their perception: two different users can have different ideas regarding the category to which a certain document belongs. Another problem that can arise is that when submitting documents to a repository, users are often presented with too many choices of categorisation, which in turn lenghtens the content posting process, introduces more inconsistency, and higher costs.

c. **Parsing and natural language analysis:** Much effort has been put into an obvious approach to deal with unstructured information, which is parsing (also called semantic or lexical analysis). Here rules of grammar and lexicons are applied to try to explicitly understand textual information. In spite of more than 20 years of researc into parsing, little inroads into real applications have been made, mainly because of its poor performance for real world problems. This method offers lack of context, unreliability, language-dependency and has a requirement for manual labour.

4. Conclusion

iLevel Software’s best of breed enterprise software solutions provide a scalable and cost-efficient alternative to existing approaches in managing the entire life cycle of unstructured and structured data. With the tagging and categorisation engine, iLevel Software offers a proven and far superior approach to the retrieval of unstructured information when compared to other alternatives and solutions. Organisations in all sectors can boost productivity and savings by setting up a unified knowledge access point, with content instantly accessible via a central portal, a Yahoo!-style Directory of topics, or a powerful taxonomy-based search and retrieval engine. Business authors can collaboratively manage and post information, while end users are able to quickly discover exactly the data they need, in context.

More reading material:

> http://www.ilevelsoftware.com/products/factsheet.pdf
Insite Server product factsheet

> http://www.ilevelsoftware.com/products/InsiteServerEDM.pdf
Insite Server Enterprise Document Management

> http://www.ilevelsoftware.com/products/Insite Server - XML Native Management.pdf
Insite Server native XML management

Contact information

United Kingdom
iLevel Software
Innovation Labs, Watford Rd
Harrow Greater London, UK
HA1 3TP

Tel: +44 (0) 20 8357 7344
Email: uksales@ilevelsoftware.com

USA
iLevel Software
245 Park Avenue, 39th Floor
New York, 10167 USA

Tel: +1 212 672 1820
Email: ussales@ilevelsoftware.com

About iLevel Software:

iLevel Software offers software solutions that elevate standards for collaborative, browser-based management of information. iLevel Software's widely used Content and Document Management Software solutions are unmatched in features, scalability and flexibility. iLevel Software has offices in London and New York, and has a growing number of partnerships with established software resellers (VARs), consultants and ASPs throughout Europe and North America.