Automatic Generation of Meta Tags for Intra-Semantic-Web

Dr. Damir Ćavar and Dr. Uta Störl[*]
Dresdner Bank AG
IS-STA Software-Technologie und –Architektur
für Allianz-Gruppe Deutschland
Research and Innovations
D-60301 Frankfurt am Main
{Damir.Cavar, Uta.Stoerl}@Dresdner-Bank.com

Abstract

Meta information in documents is very useful for the management of documents and for information retrieval. Meta information helps in search processes where necessary information is not overtly contained in the text of the document itself, or in documents that contain non-textual information. Assigning topic or keywords to documents helps in identification of relevant documents in search and retrieval processes. However, assigning keywords to documents is a problematic task for human editors. In this paper we present a solution for seamless automatic assignment of topic and keywords to documents in different formats, integrating the automatically generated meta information into the documents themselves or generating XML- and HTML-format documents, obeying Semantic Web standards.

1 Introduction

Document and content management systems make use of meta information in documents. Most editing tools and office systems provide the necessary extensions for meta information, be it text or graphic data. No doubt, meta information is very useful for the management of documents and for information retrieval. Meta information helps in search processes where necessary information is not overtly contained in the text of the document itself, or in documents that contain non-textual information. Assigning topics or keywords to documents helps for identification of relevant documents in search and retrieval processes.

Assigning keywords and topics to documents as meta information is a very useful, but also a very painful task. On the one hand, the attempt to define keywords for a document confronts the author with cognitive effort that consumes too much time and energy. On the other hand, the assigned keywords seem to be sensitive to subjective mood, varying

[*] Many thanks to our colleagues Richard Kauppert and Martin Lorenz for their help und comments. And, many thanks to the reviewers, who not only provided good comments and hints, but also interesting ideas on alternatives and future development.

on the situation of the author or the time of the day. Often enough, such keywords tend to be too general, thus reducing their usefulness extremely.

Furthermore, a unified linguistic basis seems to be difficult to establish. Certain keywords preferably seem to appear in their plural form, while others appear as singular. The bet is that most financial documents where the topic has to do with "shares" will also receive "shares" as a keyword, rather than "share", similar with "stock options" (rather than "stock option"). On the other hand, letters to family members will be assigned keywords like "grandmother" and "grandfather", rather than the plural form of these nouns. There might be some system behind such preferences. Nevertheless, a machine-based solution might prefer a unique morphological basis for keywords, i.e. the lemma or lexical base form (e.g. nominative singular for nouns, infinite form of verbs).

A potential solution might lie in automatic keyword generation and topic detection. Reliable automatic keyword generation might generate more standardized meta tags for documents, in a unique and standardized linguistic form, e.g. lemma rather than some morphological variant of a word. Such an automatic meta-tagger reduces the time and effort for the author and might even be plugged into standard applications seamlessly.

In the following we shall describe a project for automatic meta-tagging of standard office documents and document corpora with the use of text-mining and standard linguistic components. The system is realized as a Web Service and based on a linguistic tagger and text mining components and generates e.g. RDF-conform documents with Dublin Core tags in RDF-format. In a further example implementation it is seamlessly integrated in a standard text processor and generates automatically meta tags that are added to the original document itself (Word, but also HTML- or XML-format possible).

2 Analysis of Documents for Meta tagging

Different properties of documents and text can be generated automatically. Among others, the following information can be extracted from structured and unstructured text automatically:

- Language of document, paragraph, or sentence
- Topic of document
- Specific keywords in document
- Specific keywords for one document in relation to a collection of documents

Language recognition is very robust and available either as a commercial[1] product or as Open Source. For language recognition we make use of commercial products provided by Temis Deutschland GmbH, embedding the Xelda tools (Xerox).

[1] Commercial language recognizers are available in different products from various companies, i.e. Xelda tools (Xerox). This functionality is already integrated in e.g. recent versions of Microsoft Word.

For topic detection – not the detection of topic structures, rather the detection of the general topic of the document – different technologies might be used. On the one hand, a fixed taxonomy, used as a classification matrix, can be implemented as a topic matrix as well. That is, documents are classified on the basis of a predefined taxonomy with specific features and some classification algorithm. A classification algorithm can be based on extraction of substantive words and the matching of the extracted word list with word lists organized in a taxonomy. The word list can now be used for topic assignment to documents on a basis of a fixed taxonomy. On the other hand, the extracted word list can be used as a keyword list, if the most significant keywords are chosen. In this scenario, the parameters to be set are the domain specific taxonomy and the definition of "most significant" with respect to the keyword list. This type of "topic detection" and keyword assignment is possible with different tools for document classification or categorization.

In our project we make use of a simple text frequency (tf) analysis of substantive words for simple keyword extraction. However, the keyword list is normalized by linguistic tagging and lemmatization in a first step. That is, we generate lists of lemmatized (i.e. the lexical normal form) of substantive words per document in a logarithmically scaled frequency based order. Substantives are in our scenario only nouns and nominal groups, not function words, neither verbs, nor adverbs etc. Furthermore, we provide a selection schema for the final keyword list on the basis of either the maximum count for keywords or frequency threshold. A second selection and weighting schema for the keyword list is based on the text and (inverse) document frequency (tf-idf)[2] of the extracted substantives. Here the distribution of terms over all documents is used to determine a text-specific weighting. The underlying intuition is that terms that have a more or less equal distribution across all documents, are less significant for each specific document. This weighting is (for the time being) only useful for offline annotation of corpora, as described in the use cases in section 3. Making use of such a weighting algorithm would require more complex architectures of the underlying analysis component.

An independent component is responsible for the classification of nouns and nominal groups as named entities[3]. This task is solved with the Temis Extractor (www.temis-group.com) provided by the Temis Deutschland GmbH, on the basis of finite state automata for German and English. For the time being, we focus on the recognition of organization and person names.[4]

The collocations of substantives are used to generate not only a list of significant substantives, but also the relations between these words or terms.

[2] See e.g. [MS99] for a detailed discussion of such algorithms.

[3] We use the MUC-7 specification and named entity set available at www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.

[4] For more detailed discussion of different language processing methods and tools (e.g. tagging, lemmatizer, processing with finite state methods), see e.g. [JM00].

Another method that we evaluate – with the same tools and methods, as mentioned above – is based on the extraction of correlations of terms on a linguistic basis. The simple intuition is that the terms that are collocated in syntactic (or linguistic) domains, like for example the sentence domain, are semantically related. That is, we assume that if the collocation domain is syntactically defined, i.e. restricted to the basic linguistic domain of a sentence, then the relation between the extracted pairs of substantives or terms is surely semantic in nature, since they are both arguments or modifiers of the sentence predicate, e.g. the main verb. This way we extract linguistically normalized simple term pairs that are semantically related. These term pairs undergo the same frequency analysis as the single terms mentioned above.

On the basis of the discussed technologies, which we use for example for semantic net generation,[5] we are able to extract:

- Keyword lists from single documents
- Semantic relations between elements in documents
- Language of the document
- Certain named entities

Technically, the extraction of terms from documents is organized as follows:

1. Document conversion: X-format → ASCII/Unicode
2. Language recognition
3. Tagging and lemmatizing
4. Filter and analysis

In the first step, documents are converted to some basic format. For the time being we are restricted to ASCII, however, the target format is planed to be some Unicode standard (e.g. UTF-8). The converted document is passed to a language recognizer, which defines the choice of a language specific tagger. The tagger annotates every single word linguistically and the lemmatizer finds for every word the underlying lexical form, i.e. the lemma for the respective word. The linguistically tagged document is filtered and analyzed. As mentioned earlier, finite state automata are responsible for the recognition of linguistic properties like noun phrases or named entities, and the filtering of irrelevant terms (non-substantives, i.e. function words, adjectives, adverbs, verbs). For a correlation/collocation analysis a sentencer has to limit the extraction of term-pairs on the sentence domain.

There are, of course, many ways and approaches for clustering, classification and topic detection, i.e. many purely statistic approaches serve this purpose. We concentrate on the results and evaluate both types of methods, i.e. statistical methods and pure linguistic

[5] The K-Net project of the Research & Innovations group at the Dresdner Bank AG developed a semantic net generator, basically as an add-on to standard search engines with the capability to provide a view on the term relations found in documents in the intranet. See [CA01] for more details on the K-Net project.

methods for backend task of keyword extraction. Since our focus lies on meta tagging and annotation of documents, we are not primarily interested in the underlying technology of the backend, rather in the quality of the results.

Given the basic analysis of documents described so far, we are in principle able to use machine translation tools for the generation of multilingual versions of the extracted keywords and term pairs. This task is saved for future work.

The tools described so far generate the following additional information:

- Language of document
- Significant word list (our keywords)
- Term pairs, potentially with labeled relations (our semantic net)
- Topic (via clustering and mapping to taxonomies or via aggregation of semantically related term pairs)

Given this additional information, we are able to add the meta information to the analyzed documents. The benefit is clear, because standard search engines nowadays are aware of meta information (meta tags), we expect much better results for search, processing, and for general purpose knowledge management.

The question is, how this information should be generated and how it should be stored. Looking at our K-Net project (see footnote 5) we are currently able to store the information in a dedicated database. This proprietary storage form requires special interfaces for retrieval and processing. Our aim, however, is to create a non-proprietary or standardized way to store the documents and meta information and make use of existing technology for processing and analysis of such standards.

The technology we have chosen is XML-based, with the different coding standards or dialects RDF (http://www.w3.org/RDF/), Semantic Web (http://www.w3.org/2001/sw/), Dublin Core Metadata Initiative (cf. [KS01]) and Text Encoding Initiative (http://www.tei-c.org/). We are evaluating how to integrate these coding standards into proprietary document formats like for example MS Word, in order to make conversions to formats like XML or HTML as painless as possible with respect to the meta information mentioned above.

With respect to the integration of the technology to extract meta information and make it persistent, we stick to the principle of “minimal invasive knowledge management”, as discussed in [CK02] and [Ca02], i.e. the technology is seamlessly integrated into standard tools and existing infrastructure, without consequences, on the one hand, for the habits and practice of the users and, on the other hand, without consequences and changes of business processes. “Minimal Invasive Knowledge Management” refers to unobtrusive technology for knowledge processing and representation.

In the following sections, different variants of automatic meta tagging and document annotation are discussed in more detail.

3 Meta Tagging Documents

After generation of meta information or further information about the content, this information has to be added to the document and managed appropriately. One has to take into account that business documents typically are available in different proprietary formats, like for example MS Word or PDF, rather than XML. Approaches to use some XML format as a general document format, as realized in recent OpenOffice and StarOffice, are rather exceptional [SD01]. The challenge thus is to cope with the different document formats and add the new information to the documents regardless of their basic format. There are numerous technical solutions for this problem. Three solutions will be discussed in the following.

3.1 Variant 1 – Create additional information about documents

The first variant is the creation of additional information about documents, storing the additional information in related independent files. Each document is related to one XML-document that contains the generated metadata and a link to the original document. The link format (HTML-reference, XLink format etc.) depends on the search engine that is being used and makes use of such links. Such XML-documents look like:

```
<?xml version="1.0" encoding="UTF-8">
<document>
<rdf:RDF xmlns:rdf=...>
...
</rdf:RDF>
<document_link>
  Pressemitteilung0203.doc
</document_link>
</document>
```

Figure 1: Example Variant 1

The advantage of this variant is that the respective documents do not have to be converted and thus all document types that are supported by the analyzing components can be handled. However, certain mechanisms have to be implemented in order to ensure the consistence of information, i.e. regular checking for changes and existence of the documents and change of the respective metadata. A disadvantage of such a solution is that only metadata can be added to the documents. Additional XML-tags for example for named entities can only be added to the original document.

3.2 Variant 2 – Converting Documents to XML

In order to add further information to documents it is possible to convert the documents to a different format, e.g. XML or HTML. There is no universal conversion tool for all document formats. The different document formats require specific conversion tools. One way is to convert all documents to ASCII. Conversion to ASCII, however, results in loss of the documents structure or semantics, if structural information like paragraph and section disappear or title and author are eliminated in the conversion process.

Most Office Suites offer some interface to specific conversion functionalities. Microsoft Office documents can be converted with remote access over the COM interface, while OpenOffice or StarOffice 6.x documents are already stored in XML-format. The Postscript format or the Adobe Acrobat PDF can be converted with the use of free and Open Source software. After conversion to ASCII (or some Unicode format) the necessary XML-header is generated and merged with the respective meta information, a link to the original document and the document text. The text of the document is embedded in a <document_body> ... </document_body> tag. The pure text can be extended with specific tags annotating named entities, if the necessary text mining tools are available. Alternatively, the respective Information about a document can be stored in a database.

```
<?xml version="1.0" encoding="UTF-8">
<document>
<rdf:RDF xmlns:rdf=...>
...
</rdf:RDF>
<document_link>
  Pressemitteilung0203.doc
</document_link>
<document_body>
Die Geschäftsergebnisse der <company>Dresdner Bank</company> im Jahr
<year>2001</year> ...
...
</document_body>
</document>
```

Figure 2: Example Variant 2

Additional information about the language and even different language versions of the same document can be created and stored in the XML-mirror as well.

Depending on the document type, all kinds of different information might be available for extraction and annotation within an XML-format. Consequent use of style sheets for office suits can facilitate the annotation of structural and semantic properties of text (sections, paragraphs, tables, images etc.), extending the possibilities for search engines. Commercial tools are available for such a purpose. Most office suits offer some macro language and interface for extensions.

Depending on the number of different document types to be supported, the document specific conversion utilities imply some effort. The additional memory for the parallel XML-version mirror of the document basis doesn‘t appear to be that relevant, since usually the original formats of certain office suits are much bigger. The problem with the actuality and consistence of the XML-mirror still exists, analogues to variant 1 above.

3.3 Variant 3 – Storing Meta Information in the Documents

In a third variant the information is stored directly in the document itself, in the specific format. Many file formats provide support for meta information, which is already exploited by different search engines and extraction tools, e.g. PDF, MS Word, StarOffice. Certain document formats allow for a free list of key-value pairs to be defined in a document, e.g. MS Word. This possibility allows for specific setting of RDF and/or Dublin Core properties in the document.

In an example implementation we make use of a Web Service based keyword extractor that is connected with Microsoft Word. Via a macro in Word the content of the present document is sent to the extractor that returns a keyword array ordered on the basis of weights. As mentioned earlier, the weighting is based on a text frequency analysis of the linguistically normalized terms. Either the maximal count of keywords or the weight threshold can be used to restrict the size of the keyword array. In a first implementation the macro is linked to the saving button of the text processor. The resulting keyword list is automatically added to the properties of the active document.

The advantage of this variant is that no redundant information is created and thus no maintenance of consistency has to be organized. The disadvantage of this variant is that the integration of the respective tools and the machinery for handling the additional tags requires specific development for each document type and text processing system. Furthermore, the specific information must be made available for search engines. Integration of additional information in the text directly, by annotation of certain text elements is problematic, rather impossible. Here we depend on the specific text processing system and its ability to cope with such annotation.

With this option different scenarios are possible. Different technologies can be integrated directly in the text processing system, seamlessly doing the work in the background. Documents can be annotated and meta tagged without active interaction with the user.

3.4 Comparing the Variants

The following figure illustrates the ideal core conversion and annotation workflow of the meta tagging system:

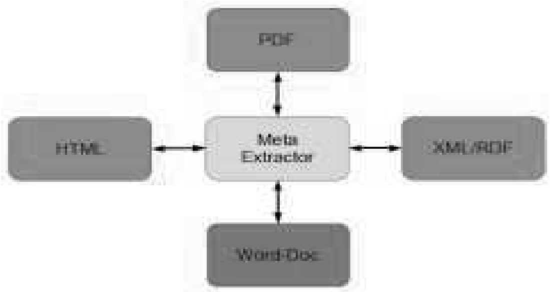

Figure 3: Conversion and Annotation Workflow

In the ideal scenario, any document format can be annotated and transformed into any other format listed in the figure above. However, for the time being we restrict ourselves to a solution, where a proprietary format (MS Word) can be annotated and converted to the other standardized formats.

The architecture of the meta tagging system is the same for all the variants discussed above. The difference lies in the output of the conversion and analysis tools. An overview of the architecture is shown in the following figure:

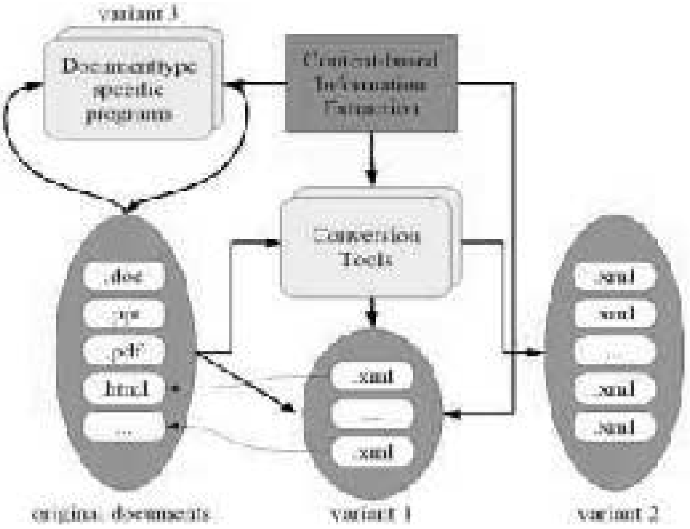

Figure 4: Different Conversion and Annotation Variants

The advantages and disadvantages of the different variants are summarized in the following table:

	Redundancy	Document type specific effort	XML-annotation in documents
Variant 1	Yes	Little	No
Variant 2	Yes	Medium	Yes
Variant 3	No	High	Partially

Table 1: Comparing the Variants

A global conclusion on the preferable variant cannot be given. The decision for one variant depends on the used document types and many other conditions. In an environment with very homogenous document types it might be an option to invest in the document specific adoption as mentioned for variant 2 and 3. On the other hand, a very heterogeneous environment suggests variant 1. The type of document storage also influences the decision for an integrated approach as in variant 3, or a more redundant approach with the corresponding problems, as in variant 1 and 2. If the documents are stored on a central document server with a corresponding document management system (DMS), changes are controlled by the DMS, and thus the redundant XML copies might not be that problematic. If the documents are kept in group or project folders, it might be an option to store the necessary meta information directly in the documents and avoid synchronization with mirror files. On the other hand there may be legal reasons to not allow users or systems to change existing documents and consequently you have to choose variant 1.

4 Conclusion

The need for structured meta information describing (structured and unstructured) documents drives different technological developments, i.e. standards and tools for information processing like XML and search engines. A fundamental problem coming with these developments is the additional effort required to efficiently leverage this technology. Even worse, the additional effort is most often perceived by the author or producer of information and not the user of the new technologies.

Our focus lies on the development and integration of technology that relieves the information producer from the duty to provide structured and also redundant information about documents. We are strongly convinced that the architecture and integration of such technology has to follow certain principles that can be summarized in the following points:

- Use and integration of existing technology platforms
- Avoid proprietary solutions
- Least effort principle from the user perspective
- Minimal invasive technology

As mentioned earlier, a technology that obeys these principles in knowledge management (KM) infrastructure we describe as “minimal invasive knowledge management” (cf. [CK02]). Our Intranet represents one component of the KM infrastructure of the Dresdner Bank. Our aim is to raise the quality of this KM-component with the use of automatic annotation and Semantic Web standards, thus creating a “Semantic-Web-Intranet”. We consider the Semantic Web related projects, the XML standard and meta tagging an important integral part of a KM infrastructure.

The way we make use of text mining and linguistic tools has one further advantage. So far we are completely domain independent. That is, our analysis components do not require textual data from certain domains, e.g. financial or banking domain. While there are certainly many ways to extend or tune the system, for example with the generation of keywords not contained in the documents on the basis of e.g. mono- or multilingual thesauri[6], we prefer to stay domain independent. Thesauri are usually either domain specific or useless. We consider the use of domain specific thesauri rather as client side

[6] This interesting idea was brought to us by one anonymous reviewer.

solution, via query expansion tools, that respect some choice made by the user or refer to some user specific domain.

With respect to the basic problems of Semantic Web related technology, i.e. the effort that is needed to create a basic knowledge space of value, we expect from tools for text mining and linguistic analysis to solve part of the problem now. Most probably, in the near future we will be able to extend automatic annotation with much better methods and algorithms. Both types of technology together make sense. The success of such technology will depend on the principles above. Our meta tagging project is a proof of concept for both, the technology and these principles.

References

[Ca01] Ćavar, D.: "Natürlichsprachliche Systeme: Natürlichsprachliche und wissensbasierte Systeme im Finanzsektor." *Focus on Research 5*, Dresdner Bank AG, STA Research Papers. 2001.

[Ca02] Ćavar, D.: "Generierung von Semantischen Netzen für das Wissensmanagement." Paper to be presented at the 9. AIK-Symposium "Semantic Web" in Karlsruhe. April, 2002.

[CK02] Ćavar, D.; Kauppert, R.: "Strategien für die Implementierung von IT-basierten KM-Lösungen: Minimal Invasive Systeme." Focus on Research 6, Dresdner Bank AG, STA Research Papers. To appear in: C. Prange (ed.) *Organisationales Lernen und Wissensmanagement: Praxiserfahrungen aus Industrie und Beratung*. Gabler Verlag, 2002.

[JM00] Jurafsky, D.; Martin, J.H.: *Speech and Language Processing*. Prentice Hall, 2000.

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

[KS01] Kokkelink, S.; Schwänzl, R.: Expressing Qualified Dublin Core in RDF/XML. http://dublincore.org/documents/2001/08/29/dcq-rdf-xml/. 2001.

[SD01] Störl, U.; Deppisch, U.: XML-basierte Content-Analyse: Vision und Realität. Pages 485-490 of: *Datenbanksysteme in Büro, Technik und Wissenschaft* (BTW), 9. GI-Fachtagung, Oldenburg. Springer, 2001.