Exploiting Metadata for Ontology–Based Visual Exploration of Weakly Structured Text Documents

Christian Seeling, Andreas Becks
Fraunhofer Institute for Applied Information Technology FIT
Schloss Birlinghoven
53754 Sankt Augustin, Germany
Tel: +49 241 80 21503
{christian.seeling,andreas.becks}@fit.fraunhofer.de

Abstract

A large amount of strategically relevant business information is contained in unstructured texts. While information brokering approaches are used to contextualize such documents and to generate metadata, text mining is used to explore large document spaces. So far, little attention has been paid on a value-adding combination of these technologies. In this paper we show how metadata and documents can be complementarily represented and used interactively to support users in text corpus analysis. We present a text analysis portal which displays inter-document similarity by means of so-called document maps, complemented by a display of the domain ontology and metadata-based access methods.

1. Introduction

The metadata description standard XML, introduced by W3C in November 1996 and revised continuously ever since, has contributed significantly to the increase of investigations in metadata descriptions in research and industry. The semantic web and the use of ontologies open up new potentials for processing semi-structured information resources.

In particular, business applications may significantly benefit from these developments: Today, managers are often confronted with the fact that most of the business-relevant information is neither structured, nor annotated or at least formatted in a unified way. Mayer and Freiberg [12] state that by far most of the strategically relevant information is encoded in natural language. Management reports, surveys and news tickers are only some examples. At the same time, Uhr [16] emphasizes the importance of external information for the adequate evaluation of company performance and for planning tasks in the light of global markets.

The dilemma of missing accessibility to crucial information (caused by the natural language encoding and missing unification in format and structure) can be addressed in two ways: On the one hand, *information brokers* who mediate between information providers and consumers [14] offer the service of annotation, classification, and personalization of information, playing the role of metadata-generators. On the other hand, the field of *text mining* attempts to cope with unstructured textual information by merging elements of information retrieval and explorative text access, yielding a powerful method set for text analysis [9].

Right now, metadata is mainly used 'in the background' – hidden to the user – as a backbone for searching, extracting, integrating and reasoning about information. However, making metadata directly available to users promises to maximize its utility: Consider for instance applications in financial planning. Here, metadata can be used to mark-up important business data in texts (e.g. business measures), helping managers to analyze information on the market situation. But also the unstructured parts of that information (like business stories) are mission-critical, since they convey a picture of the whole situation.

In this work we introduce a method that combines metadata-based information brokering and text-mining in order to help users to exploit both types of information. A text analysis portal is presented where metadata descriptions of documents (called 'contexts' according to [11]) and inter-document similarity are visualized, enabling the analyst to simultaneously examine documents on a conceptual and natural-language level. Both, navigation and metadata representation are based on a brokering domain-ontology. Based on a careful analysis of requirements (section 2) and a discussion of related work (section 3) we present design and architecture of the portal and show how practically-relevant text corpus analysis tasks can be performed (section 4). A discussion of application experiences (section 5) completes this paper.

2. Metadata-Based Text Corpus Analysis

Information Brokering makes several information processing steps necessary. Figure 1 illustrates the contextualisation and the personalisation according to models of domain and user. We focus on the support of the third role (besides the roles of information *provider* and *mediator*),

namely the information *consumer*, i.e. the analyst who is interested to examine the delivered information by searching, exploring and structuring it.

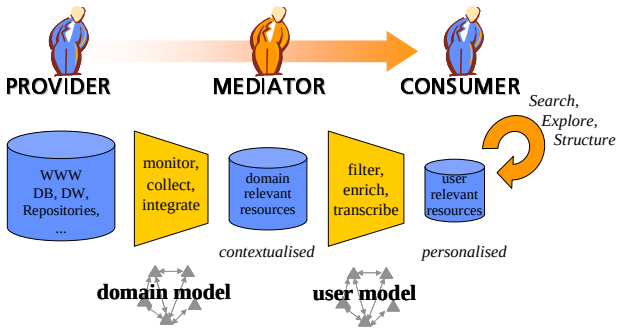

Figure 1: Integrating Information Brokering and Text Mining: Information Flows

In the remainder of this chapter we describe an exemplary scenario of text analysis to explain typical questions and needs. Then we derive requirements for a befitting technological support.

2.1. Analysis of Market Actor Profiles

Market actor analysis is an important instrument for the very early planning of upcoming production lines as well as for the continuous monitoring of partners and competitors. Analysts as well as business players are often interested to get a compact overview of actors in a specific market segment. For illustration we address companies in the textile production and trading industry, a field into that we recently could gain deeper insight during a current research project [3]. A market analyst is interested to answer questions like the following:

- Are there key actors / monopolists / niches?
- Which groups of companies do exist with respect to product, sub-sector, size, turnover, customer-group, production capacities, region, etc?
- How do the criteria depend on each other? For instance: Do traders of sports wear have a typical legal form? Which influence factors are correlated with the company size?
- Which buzzwords and arguments are commonly used in corporate identities?

There are many providers[1] offering collections of company profiles that are typically semi-structured according to predefined templates. Such web repositories can be accessed by searching the documents and metadata or by browsing through taxonomies.

However, the fine-granular examination of company profiles, including the examination of relations between different companies, is an unsupported task of the analyst. There is a lack of support considering the compactness of the access interfaces and contrariwise the high complexity of the analysis requirements.

2.2. Requirements

As a guiding principle we demand that the available metadata is exploited as far as possible to allow the combined analysis on metadata level and document level. Especially the contrast of explicitly modeled, structured metadata on the one hand and weakly-structured text information seems to be valuable since it can generate hints by raising questions (How are these classes related with respect to the text collection? How are these documents that share many important terms characterized in the model of the domain?). This effect can only be reached if both aspects are equally weighted what concerns the richness of functionality and the fraction of the user interface. Furthermore, both views should be tightly integrated so that each navigational step in one view is simultaneously presented in the complementary view. To prevent the analyst from being overwhelmed with information, he should always be given an overview of both, the metadata model (domain ontology) and the document space.

What concerns functional requirements, there are three core analysis activities that should be supported: For the support of exploration, the *characterization* of both, groups of documents and sets of metadata, is crucial. A system should also provide a means of supply for *navigation* based on the metadata model. For the support of *retrieval*, it is consequent due to the mentioned paradigm of tight integration to not only offer the search in unstructured text but also the ontology-driven search.

Due to the fact that visual perception is a human key skill (cf. M. Hearst in [1], chapter 10.2.2) and the mentioned demand for overview, a complex analysis portal should generally make use of compressing graphical metaphors for text access.

3. Related Work

In the field of database research, structured relational or object-oriented query languages are discussed [5]. Information retrieval complementarily addresses the search in unstructured information repositories. For the successful formulation of a query a good knowledge of the terminology of the text collection is critical (vocabulary problem, cf. [7]). Taxonomies offer a classification that can be exhausted for exploration. But here the lack of overview is problematic and the common terminological understanding of taxonomy-designer and information analyst is a necessary precondition for success. As was sketched in section 2.1, the use of standard query interfaces or cata-

[1] for example refer to http://www.bizzcontact.com, http://www.usawear.org, http://www.tessilmoda.com

logues does not tap the full potential of text analysis in the presence of metadata. Often, query interface and catalogues are provided separately.

There exist many prototypical and commercial realizations of information visualization systems for metadata or documents. Comprehensive discussions of potentials, drawbacks, metaphors and task-adequacy can be found in [13] and [6]. Since this work is mainly concerned with the *combined* representation of documents and metadata, we selected and criticize two representative approaches. Both systems are designed to support metadata-based analysis of text-repositories. The first (graphically aided) tool offers navigation based on a preexisting classification, whereas the second tool allows for the generation of metadata that is subsequently provided for taxonomy-based exploration. Important comparison criteria are worked out and applied.

SPECTACLE from Aidministrator (cf. [8], chapter 3) is a text corpus exploration tool mapping classes and documents to the same graphical display. Another frame shows the predefined classes that are organized hierarchically – the authors call it a 'light-weight ontology'. This ontology is used as well for the configuration of the display as for navigational purposes. Documents are classified and groups of equally classified documents are spatially grouped in document clusters. Interactions comprise the configuration of visualized classes, document details, ontology-based interactive query formulation and diverse navigation capabilities. With K2 ENTERPRISE, Verity Inc. (www.verity.com) offers an advanced information retrieval and text mining application. The extraction of important keywords serves as a basis for fuzzy searching. The software also comprises methods for concept detection and query-by-example. Most important, hierarchical categorization of documents is performed automatically. The system creates a structured representation of dominant terminological concepts, which allows for subsequent taxonomy-based document analysis. The model can be improved or refined by the user.

In both systems, the domain model is tree-like. In Spectacle class taxonomies and classifying assignments are defined manually whereas K2 employs concept extraction algorithms. Both tools adopt specialization hierarchies and overlapping classes. The Spectacle cluster visualization is well-suited for human visual capabilities but there are some critical points: It does not give any information about the similarity of documents. Often, the analyst likes to see documents related to the one or the many that he has already identified to be interesting. Furthermore, with increasing number of documents and classes, significantly difficult to keep the overview. The Spectacle tool strives for a consequent integration of metadata and document analysis: Visualization as well as navigation and querying are possible for both worlds. Nevertheless, the capabilities for document analysis are subordinate.

4. A Novel Information Analysis Portal

In chapter 2 we worked out important requirements for the metadata-based analysis of text corpora. In chapter 3 argued why existing approaches do not fully satisfy the requirements. Here we propose a new analysis portal that consequently supports metadata-based text analysis.

4.1. System Architecture

An information brokering component retrieves and filters documents according to the analyst's interest profile. The delivered information stems from web documents that are metadata-enriched (i.e. contextualised) and personalized (cf. figure 1). A document visualization and mining tool aids in examining the content of natural-language document repositories. The portal integrates both tools and offers additional analysis features. The system described is called SWAP*it* (Semantic Web Analysis Portal for intelligent text analysis). The integration is web-based using Java Applet and SOAP communication technology.

In [4] we have described a document map system for visually aiding text corpus analysis tasks in knowledge management that is called *DocMINER* (Document Maps for INformation Elicitation and Retrieval). The document landscapes that are computed fully-automatically adopt a $2^{1/2}$D geo-spatial metaphor for the visualization of inter-document similarity: The geographic distance of document symbols corresponds to the dissimilarity of the documents contents. We developed a domain-specific task model [2] that did not only serve as a guide for the technological development but also as a yardstick for evaluation. DocMINER supports an adaptable framework for generating a graphical corpus overview. Its interface design was guided by Shneiderman's 'Visual Information Seeking Mantra' [15]: *Overview first, zoom and filter, then details-on-demand*. System features include different zoom, scaling and sub-map functionality, means to define and assign document symbols, an annotation function, automatic map labeling and document group summaries, and a tight coupling with a query-driven retrieval interface. The document similarity landscape serves as a text mining workspace. DocMINER is further used to provide term statistics and retrieval functionality.

Nick and Klemke introduce their information-brokering suite *Broker's Lounge* [11][14]. The system provides a meta-model for the design of domain-specific ontologies Basic functionalities comprise modeling of domain, user interest and information sources, regularly crawling web resources, and information filtering functionality. The system is used to deliver a model of the domain as well as contextualized and personalized documents for the analysis in SWAPit.

According to Gruber, '*an ontology is a specification of a conceptualization*' [10]. Our metadata model for document description extends the mentioned model and offers the following modeling elements: A document can be assigned to one or many *concepts*. Concepts can be described by *basic-value attributes* as well as by *hierarchically classifying attributes*. A *concept pattern* consists of a concept name and concept descriptions comprising values for the concept's basic-attributes and a binary-valued classification along the concept's hierarchical classification-attributes. *Relations* among concepts can be used for information characterization. We will refer to a *document context* as a set of concept patterns and a set of inter-concept pattern relations that hold for the document. Modelling elements can be assigned to word fields ('terminological clouds'). The modeling formalism reflects our document-centric view: Documents are the 'smallest unit' for analysis. We do not expect a document to be unique or homogeneous with respect to the domain ontology. In the opposite, we intend to allow a flexible metadata-description of documents. Consequently, documents are not seen as instances of concepts but are rather embedded in a metadata document context. The formalism does pragmatically link networks of concepts (known from ontologies and database research) with classification trees, often used for text mining and text access (and widely known from file systems and web catalogues).

4.2. User Interface and Functionality

The user interface is divided into four parts as illustrated in figure 2: The upper left-side panel visualizes the inter-document similarity (computed from documents or metadata). The user can switch between a detailed view of selected documents and a general view where he can see how the selected documents are embedded in the document repository. The lower left-side panel shows a part of the domain ontology that serves as an alternative workspace for metadata-based navigation. The upper right-side panel simply displays URLs and metadata of all documents. The lower right-side panel offers multi-purpose analysis features (fulltext, search, statistics)

To make the contrast between metadata document context and document fulltext explorable (cf. section 2.2), two separate selections of documents and classes are necessary (these are colour-coded). When a user interacts with the hierarchical metadata model, filters are immediately applied to select and highlight documents in the document similarity workspace. Tree-based navigation we offer different options for the evaluation of multi-selections (union / intersection of assigned documents).

Additionally, the system offers a matching measure for the deeper characterization of the documents that match the current selection: The better a document is characterized by the current selection in the domain ontology, the more intense is its red colour (vice versa is similar: The better a class describes the group of selected documents the more intense is its yellow colour). This feature is similar ranking concepts that are accepted in IR.

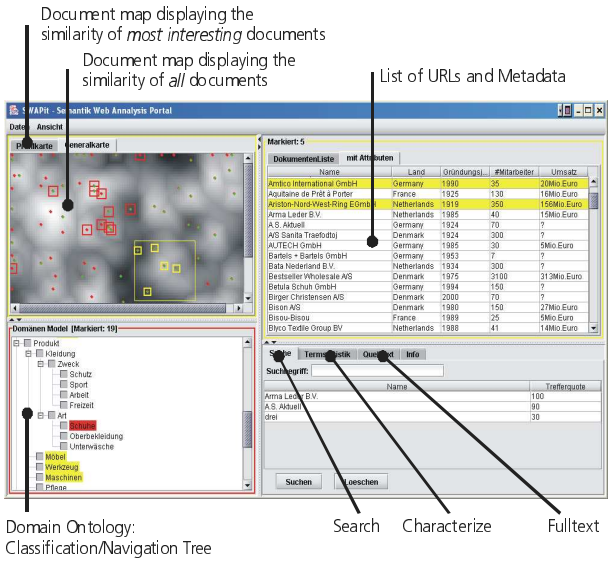

Figure 2: Interactive Features of the Ontology-Based Analysis Portal

4.3. Discussion of the system

The SWAPit analysis portal is a new medium that integrates the perspectives of text analysis and metadata analysis and offers novel analysis capabilities like the 'roundtrip' (i.e. swapping between the two workspaces and swapping back and so on) which allows dynamically contrasting document context and documenting content. The user is supported to make extensive use of the available metadata since not only navigation is based on the domain ontology but also the values of structured numerical or textual metadata can be analyzed by browsing, characterization and search. In the user interface, the extent of functional support as well as the fraction of the display of metadata and document are taken into account *equally*. The overview is granted since both workspaces (document map and ontology navigation tree) are always visible in parallel and updated simultaneously on changes. The focal subset of interest can be inspected in detail or explored in relation to the rest of the collection. We decided not to mix up the representation of classes and documents in the same graphical display to avoid confusion but we rather put emphasis on the inter-document relations. Naturally, the document contexts can be examined by interactively.

The three core analysis activities demanded in section 2.2 (characterization, retrieval and navigation) are sup-

ported: For the *characterization* of document groups SWAPit provides term statistics. In addition to these natural language-oriented descriptions a selected group of documents is described by the concept pattern (visible in the ontology tree and the attribute-value table). The analyst can use Boolean *search*, interactively configure a constraint on the domain ontology (by selecting classes in the classification tree) or group and search metadata values in the respective panel. *Navigation* on the document map comprises the selection of document groups. A click on a single document opens the fulltext-view and shows the document's context in terms of domain ontology and metadata. Furthermore, search results and the results of term statistics can be explored on the map. Navigation on the navigation tree consists of the selection of one or many classes at different levels in the tree.

SWAPit combines a document map and a class taxonomy. The early design of the system was revised and improved in several discussion-modification cycles accompanied by experts from the usability lab of our research institute. Strength of the approach is its flexibility: The approach can be generalized from the analysis of documents to the analysis of more complex objects like interest profiles. The modular design of SWAPit and interfaces allow the exchange of subcomponents: One could for instance imagine exchanging the information brokering component by a conceptual semantic web search engine. SWAPit's applet architecture allows flexible integration into more comprehensive web-based information portals.

For fine-granular text analysis, the simultaneous representation of thousands of documents does not make sense. Within these natural borders, the approach is scalable with respect to the number documents and size of the metadata model.

5. Application Experiences

We describe a use case that is concerned with the analysis of market actors in the textile industry by the metadata-based analysis of company profiles. We searched a business forum (www.bizzcontact.com) for: "C*ompanies with English Portrait* AND *Textile Industry & Fashion* AND *All Countries* AND *All Regions*" and received 126 semi-structured company profiles. We describe a fine-granular analysis of the collection using a prepared domain ontology and interest profile (interest is focused on SMEs)

First of all we analyze the market on a rough level. By inspecting the document map, we detect that the SMEs strongly tend to be placed on the left-hand side of the map. To examine this observation, we select many of the documents on the document map. Now we look at the characterization by the selected group of documents in terms of its classification in the domain ontology tree. One of the classes with the highest colour intensity is 'product->garments'. This triggers the hypothesis that most SMEs are working in the garments production sector which could be investigated interactively.

Now we explore the 'product' class hierarchy in more detail since it is a natural differentiating criterion for the companies. We click on the classes one-by-one. By the size of the highlighted document groups we get an impression of how many companies are competing within each sub-sector. We notice that the shoe cluster (12 doc.) is especially homogeneous on the document map display (cf. figure 3; darker document points represent documents that are filtered with respect to the SME interest specification) and decide to investigate it in more detail.

First we perform a 'roundtrip', that is, we explicitly select the red documents on the document map (the document points become yellow) and then look again on the domain tree for a characterization of the group in terms of the ontology. We find that the companies are in deed very homogeneous (offering footwear / accessories; located in Europe). Now, we examine the outliers on the document map, i.e. 'shoe' documents that are located on the map distinct from the group. We select the company profile that is located in the lower left corner (cf. figure 3). Inspection of the domain model tells us that the company is acting in the 'work/protection' clothing sector. We examine if the other documents in the neighbourhood of that document also have to do with protection or work and find out that most of the 'work/protection'-categorized documents are in that region. This explains why the profile is not located next to the other profiles of shoe-producing companies that have nothing to do with work or protection.

To work out the contrast between the textual description and the metadata context of documents, we perform a Boolean search for '*shoe or footwear*' and get 26 documents highlighted in yellow. Now we can relate both selections aided by the document map (documents assigned to the class 'shoe' are red). We focus one company profile that is only found by the search engine but not classified with the product 'shoe'. We use the term statistics to characterize the document: Among the most important terms we find 'coat' and 'fetish'. Now we take a look at the attribute-values of that company (operating in France, founded 1925, 130 employees, turnover 16 Mil.€). For a deeper understanding we look at the classification of the document. It is assigned to 'product>>garments' and 'customer>> woman'. We finally read the document fulltext and find that the company is offering ladies outwear. In their company description they formulate the expectation that that a coat fetish trend will follow the common shoe and hat fetishes. This is maybe an interesting information for a producer in the shoe market. It is not likely that this information would have been found by solely analyzing with the help of the ontology model.

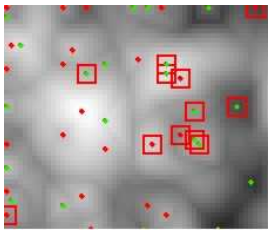

Figure 3: The Cluster of Companies Working in the Field of Shoe Production & Trading

6. Conclusion and Future Work

We stated that often information selection processes (e.g. information brokering, retrieval, querying) and information analysis processes (text mining, text access) are separated from each other. We introduced a coupling of both worlds and argued that it is important to support an analyst with metadata from the information broker. We worked out requirements and introduced a system design. The SWAPit portal integrates an information brokering and a visual text mining tool. We showed by the example of market actor analysis how the tool can be applied to generate benefits for text analysis in real-world settings.

The analysis sketched in chapter 5 points out that in spite of advanced tool support, text analysis remains an intelligible and complex task that is dependent on the cleverness and experience of the analyst. It seems to be most difficult to identify the adequate analysis strategy and to coordinate the interaction steps accordingly. Analysis processes need to be further studied and characterized (e.g. by task models like the one described in [2]). Accordingly, there is a strong need for the evaluation of the task-adequacy and the usability of supporting systems.

We are on the way to learn from several case studies, applying different hierarchical schemes for navigation (file system structures, ACM paper classification, IPC patent classification, internet-catalogues). Furthermore we intend to visualize the similarity of mixed objects (i.e. consisting of text parts and numerical metadata). To do so, we need to develop a similarity measure that can be configured flexibly for different applications. Since future semantic web search engines are expected to provide semantically rich information, our approach will contribute to make metadata accessible for navigation and analysis.

7. Acknowledgements

The authors would like to thank Reinhard Oppermann and Britta Hofmann for constructive discussions on the user interface design of the system.

8. References

[1] Baeza-Yates R., B. Ribeiro-Neto: *Modern Information Retrieval*, Addison-Wesley, 1999.

[2] Becks, A., C. Seeling: *A Task-Model for Text Corpus Analysis in Knowledge Management*. Proc. of UM-2001 Workshop on User Modeling, Machine Learning and Information Retrieval, 8th Int. Conf. on User Modeling, Sonthofen (Germany), July 2001.

[3] Becks A., Jarke M.: *Integration von Geschäftskennzahlen und Textdokumenten mit Semantic Web Technologien*, to appear in KI, Heft 4/2003.

[4] Becks A.: *Visual knowledge management with adaptable document maps*. GMD research series, 2001, 15.

[5] Bonifati A., Ceri S.: *Comparative Analysis of Five {XML} Query Languages*. SIGMOD Record vol. 29/1, pp.68-79, 2000.

[6] Däßler R.: *Informationsvisualisierung – Stand, Kritik und Perspektiven*. In: Methoden / Strategien der Visualisierung in Medien, Wissenschaft und Kunst Wissenschaftlicher Verlag Trier (WVT).

[7] Furnas G.W., T.K. Landauer, L.M. Gomez, S.T. Dumais: *The Vocabulary Problem in Human-System Communication*. Communications of the ACM, vol.30/11, pp.964-971, 1987.

[8] Geroimenko V., C. Chen: *Visualizing the Semantic Web – XML-based Internet and Information Visualization*; Springer Verlag, 2002.

[9] Gerstl P., M. Hertweck, B. Kuhn: *Text Mining: Grundlagen, Verfahren und Anwendungen*; HMD Heft 222, S. 38ff, Dezember 2001.

[10] Gruber T.R., *A translation approach to portable ontologies*. Knowledge Acquisition, 5(2):199-220, 1993

[11] Klemke R.: *Modelling Context in Information Brokering Processes*. Doctoral Thesis, RWTH Aachen 2002.

[12] Mayer R., Freiberg U.: *Workmanagement – eine wichtigeBasis für Wissensmanagement*; Jan. 2000; Zeitschrift „Wissensmanagement“ Heft 1.

[13] Morse, E., M. Lewis, R. Korfhage, and K. Olsen. *Evaluation of text, numeric and graphical presentations for information retrieval interfaces: User preference and task performance measures*. Proc. of the IEEE International Conf. on Systems, Man, and Cybernetics, San Diego, CA, 1998, pp. 1026-1031.

[14] Nick A.: *Personalisiertes Information Brokering*. Doctoral Thesis, RWTH Aachen 2002..

[15] Shneiderman B. *The Eyes Have It – A Task by Data Type Taxonomy for Information Visualizations*. Proc. of the IEEE Symposium on Visual Languages, pp. 336-343, IEEE Computer Society, Boulder, Colorado, 1996.

[16] Uhr, W., ed: *Externe Daten in Management Support Systemen*. Special Issue, Wirtschaftsinformatik, 41(5), pp. 403-457, 1999.