Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2004 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2004

KeyTEx - An Integrated Prototype for Semi-automatic Metadata Assignment and Network-based Content Retrieval

Alexander Benlian
University of Munich

Florian Wiedemann
University of Munich

Thomas Hess
University of Munich

KeyTEx – An Integrated Prototype for Semi-automatic Metadata Assignment and Network-based Content Retrieval

Alexander Benlian
University of Munich
Institute of Information Systems and New Media
Ludwigstr. 28, 80539 Munich, Germany
benlian@bwl.uni-muenchen.de

Florian Wiedemann
University of Munich
Institute of Information Systems and New Media
Ludwigstr. 28, 80539 Munich, Germany
wiedemann@bwl.uni-muenchen.de

Thomas Hess
University of Munich
Institute of Information Systems and New Media
Ludwigstr. 28, 80539 Munich, Germany
thess@bwl.uni-muenchen.de

ABSTRACT

In print and online media companies, editors and archivists often have difficulties with assigning relevant metadata to written text in a timely manner. Additionally, they often struggle with finding adequate and useful media content in hierarchically structured archives when searching for background material. In this paper, we describe the potentials of and synergies between text mining and XML topic map technologies for more efficient metadata assignment and media content retrieval processes on the basis of the open source-based Java-prototype KeyTEx[1]. By illustrating the architecture and relevant functions of the prototype, we show how both types of technology can fit together. Finally, empirical evidence on the economic application of KeyTEx is provided by presenting first findings of a lab experiment. We conclude with assumptions on the relationship between browsing performance and underlying information topology in dependence of the given search setting.

Keywords

Text mining, metadata assignment, keyphrase extraction, XML topic maps, information retrieval, semantic networks

INTRODUCTION

In print and online media companies, editors and archivists often do not have sufficient time or the traditional skills of librarians to formulate relevant and objective metadata for produced media content (e.g. magazine or online text articles) due to the pressure of time (Adams, 2002). In addition to that, with more and more media content pouring into media companies' databases, the editorial staff is struggling with finding adequate background material (e.g. archival documents) in deep, hierarchical folder structures during the production of new media content. Reasons why navigating in hierarchical structures is tremendously time-consuming for special kinds of search tasks are among other things long click paths due to deeply-structured hierarchies and the often ambivalent and ambiguous meaning of folder names. Furthermore, media content organized in hierarchical trees often can only be found in one folder although the plot or basic themes of the media content suggests several different associations. As a consequence, hierarchical structures seem not always to be the most adequate and efficient underlying information topology for every browsing scenario.

The mentioned deficits of entirely manual-based metadata assignment processes and hierarchical information topologies give reason to explore and examine alternative techniques that provide more efficiency in dealing with media content. In this

[1] Keyphrase Extractor and Topic map Explorer

paper, we propose a conceptual model and a prototypical realization of the integration of semi-automatic metadata assignment processes and network-based media content retrieval. Although the basic principles and techniques of our propositions can be transferred to each type of content (text, picture/photo, audio and video), our work will center on *text* mining and *text* document retrieval for the reason to reduce complexity.

In the following chapter, we start with presenting the related work in text mining and information retrieval research concluding with a motivation to which area(s) this paper can make a contribution (chapter 2). Chapter 3 lays the foundation of the paper by outlining the basics of text mining technologies, XML topic maps and an input-output-model for the conceptual integration of both functionalities in the knowledge management life cycle. The main part of the paper (chapter 4) presents the design of our prototype KeyTEx that technically integrates both text mining and content retrieval functionalities. Chapter 5 demonstrates first empirical results of a lab experiment about the comparison of browsing in hierarchical and network-based information topologies. On the basis of these experimental findings, the paper concludes with assumptions on which search settings network-based navigation can be more efficient than hierarchal navigation. The development and experimental testing of the prototype represent a first cycle of an action research project that aims at taking advantage of iterative and evolutionary prototyping.

RELATED WORK

Text mining for the purpose of keyphrase generation and information retrieval on the basis of different information topologies are two current fields of research.

With regard to *keyphrase extraction* as a sub research field of text mining, research focuses on different algorithms and methods as to how exact and fast keyphrases can be extracted out of human natural language documents (e.g. Witten and Eibe, 1999). One of the most pressing problem seems to be the extraction of descriptors that best represent the core meaning of a collection of information as a basis for later information retrieval processes (Hearst, Elliot, English, Sinha, Swearingen and Yee, 2002). In the research community, the dominant approach to the automated classification of texts into predefined categories is based on machine learning techniques: a general inductive process automatically builds a classifier by learning the characteristics of the categories from a set of pre-classified documents (Sebastiani, 2002). Many researchers are addressing these and related issues in (semi-) automated text mining making considerable strides (e.g. Sebastiani, 2002; Hearst, 1999).

In terms of *information retrieval*, literature mainly concentrates on different aspects how navigation in informational environments can be optimized conceptually and technically (e.g. Baeza-Yates and Ribeiro-Neto, 1999). Specialized studies in this research area examine the relationship between the performance to navigate in information spaces (e.g. WWW) and different underlying information topologies (e.g. Batra, Bishu and Donohue, 1993; Mohageg, 1992). Consistently, researchers discovered that network-based information structures, in comparison to hierarchical ones, can be a superior option for searching activities in particular search modes (e.g. Shneiderman, 1998). Hence, more and more discussions about network-based information visualization using semantic ontologies are emerging (e.g. Fluit, Sabou and van Harmelen, 2002).

With the emergence of research topics centering on the semantic web, questions arose how knowledge resources in the chaotic Internet could be efficiently exploited and adequately represented. Although research communities have examined text or (semantic) web mining and network-based information retrieval techniques from many different angles, very few papers can be found dealing with the conjunction of both research areas for special economic application scenarios (e.g. Böhm, Heyer, Quasthoff and Wolff, 2002). Although a number of proposals have been made separately for semi-automated engineering (e.g. Maedche and Staab, 2000) and visualization of ontologies (e.g. Fluit et al., 2002), an integrated conceptual model and prototypical realization for the interplay of both research areas is still lacking. The paper at hand tries to make a contribution to closing this research gap.

Prior to the description of the architecture and functions of the prototype KeyTEx in chapter 4, a brief overview on text mining concepts and XML topic maps should be given.

BACKGROUND ON TEXT MINING AND TOPIC MAPS

Text Mining, also known as Text Data Mining (Hearst, 1997) or knowledge discovery from textual databases (Feldman and Dagan, 1995), is a promising technology for analyzing large collections of unstructured documents for the purpose of extracting interesting and non-trivial patterns of knowledge. Dealing with these documents is challenging due to the fact that text is inherently unstructured and fuzzy. So the text mining process is divided into two phases: In the first phase, the text document is transformed into an intermediate form using computer linguistic, statistical and machine learning techniques. In the second phase, different pattern recognition methods (clustering (e.g. Ester and Sander, 2000), keyphrase extraction (e.g. Witten and Eibe, 1999), summarization (e.g. Hovy and Lin, 1999) or categorization (e.g. Sebastiani, 2002)) can be applied to the intermediate form, mainly depending on the application scenario. Since the application scenario of this paper refers to the semi-automatic assignment of relevant semantic descriptors to text documents, keyphrase extraction seemed to be the most adequate pattern recognition technique.

Topic maps are an ISO standard (ISO 13250, Biezunski, Bryan and Newcomb, 1999) which has evolved out of the research activities on self-organizing maps (e.g. Kohonen, 2001) and concept-based document retrieval (e.g. Chen, Lynch, Basu and Ng, 1993) that have been established as an answer to the problem of organizing semantic interrelationships between information units in large knowledge bases. Primarily, the XML-based standard was established to handle the construction of indexes, glossaries, thesauri and tables of contents, but its applicability extends beyond that domain. As topic maps generally lay out a structured metadata vocabulary that point to separately stored information resources, topic maps also appear to be a promising standard for network-based information retrieval (Adams, 2002). In contrast to alternative (XML-based) ontology representation formats (e.g. RDF), topic maps are used to structure a clearly bounded information base and construct links from known semantics to resources ("top-down approach"), whereas RDF, for instance, is designed to support open environments and constructs pointers from existing resources to known semantics ("bottom-up approach") (Daum and Merten, 2002). Pepper summarizes the core elements of this standard - Topics, Associations and Occurrences - as the TAO of Topic Maps (Pepper, 2000). *Topics* are the main building blocks of this structure and refer to elements in the topic map that represent the subjects being referred to. The interrelationship between different topics is formalized by *associations* which represent semantic relations. *Occurrences* link topics to one or more relevant information resources, like audio, video or text files.

With regard to the process steps in a knowledge management life cycle (e.g. Markus, Majchrzak and Gasser, 2002; Lee, Kim and Yu, 2001), a central knowledge base that serves as a drain for input information and a source for output information can be considered as a conceptual link between the above mentioned techniques (see Figure 1). To each process step of the knowledge management life cycle, special methods can be assigned that support the creation, integration, combination and leveraging of knowledge. In this paper, metadata and media content (more specifically *text documents*) are the central knowledge objects of interest.

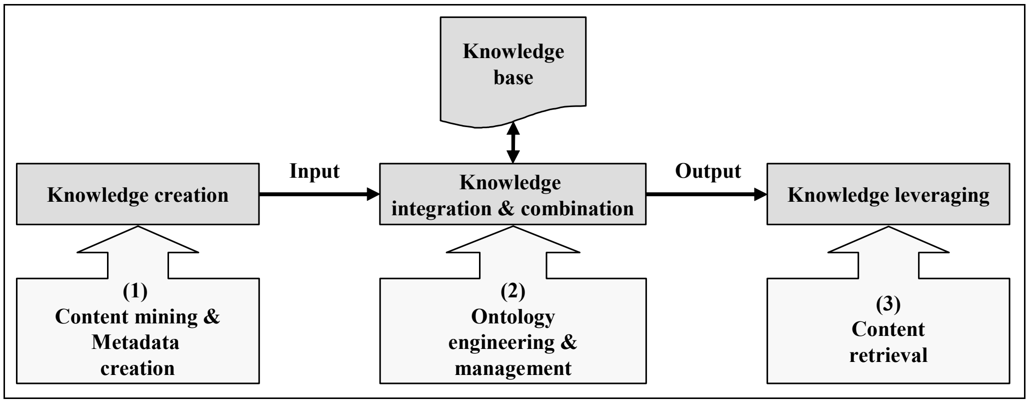

Knowledge base
Knowledge creation
Input
Knowledge integration & combination
Output
Knowledge leveraging
(1) Content mining & Metadata creation
(2) Ontology engineering & management
(3) Content retrieval

Figure 1. Input-output-model in the knowledge management life cycle

As there are many different techniques and standards within one process step in the knowledge management life cycle that can be addressed, it is important to define the scope of the presented work. Hence, a focused recapitulation of chapters 2 and 3 in the form of a morphological box seems to be appropriate (see Figure 2). The light gray boxes emphasize methods and research objects that will be addressed by the prototype KeyTEx in the different process steps of the knowledge management life cycle. For lack of space, alternative combinations of values are not further discussed.

<table>
<tr><td rowspan="2"><b>(1)
Content mining & Metadata creation</b></td><td>Mined content type</td><td colspan="3">text</td><td colspan="3">pictures / graphics</td><td colspan="3">audio</td><td colspan="3">video</td></tr>
<tr><td>Mining technique</td><td colspan="3">clustering</td><td colspan="3">(keyphrase) extraction</td><td colspan="3">summarization</td><td colspan="3">categorization</td></tr>
<tr><td rowspan="3"><b>(2)
Ontology engineering & management</b></td><td>Ontology generation method</td><td colspan="4">manually</td><td colspan="4">semi-automatic</td><td colspan="4">fully-automatic</td></tr>
<tr><td>Information topology</td><td colspan="4">linear</td><td colspan="4">hierarchical</td><td colspan="4">network-based</td></tr>
<tr><td>(XML-based) Ontology representation format</td><td colspan="3">DAML+OIL</td><td colspan="3">OWL</td><td colspan="3">RDF</td><td colspan="3">Topic maps
(XTM 1.0)</td></tr>
<tr><td><b>(3)
Content retrieval</b></td><td>Search option</td><td colspan="6">searching by browsing</td><td colspan="6">searching by (parametrically) querying</td></tr>
</table>

Figure 2. Morphological box with topics treated in this paper

THE KEYTEX PROTOTYPE

The starting point for the description of the system architecture of the prototype is a UML use case model (Booch, Rumbaugh and Jacobson, 1999) depicted in Figure 3. The Unified Modeling Language was chosen as analysis and design method, because it clearly specifies the structural and behavioral aspects of the prototype.

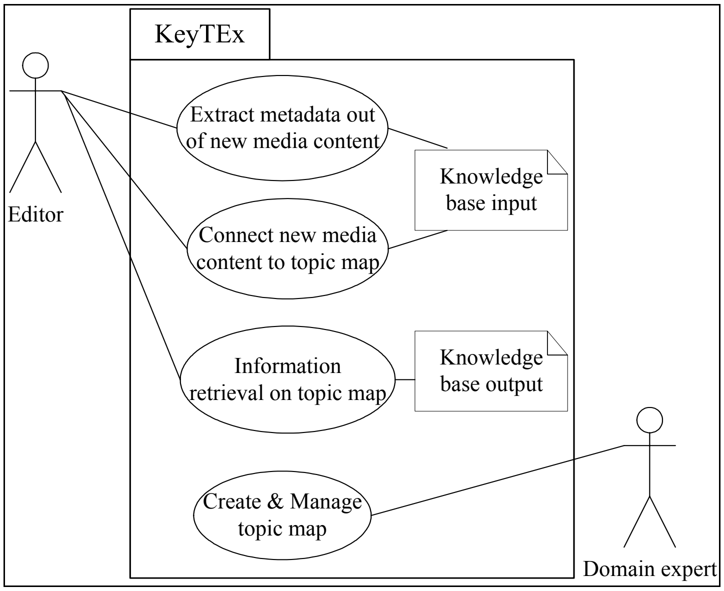

KeyTEx
Editor
Extract metadata out of new media content
Connect new media content to topic map
Knowledge base input
Information retrieval on topic map
Knowledge base output
Create & Manage topic map
Domain expert

Figure 3. Use cases for metadata assignment and network-based information retrieval

The use case model shows the basic functionalities of the integrated system provided to two different types of actors.

1. The actor "Editor" represents the role of editors and archivists in media companies that regularly need to search for archived media content and annotate newly produced media content with metadata:

- *Extract metadata out of new media content*: Editors (or archivists) usually assign metadata to text documents manually. With KeyTEx, the annotation of text documents occurs semi-automatic. KeyTEx processes newly entered or already integrated text documents, extracts a preconfigured number of keyphrases and suggests them to the editor.
- *Connect new media content to topic map (knowledge base)*: As another quality step in the annotation of media content, the editor has the possibility to adjust the suggested keyphrases manually. After the editor has performed this process step, the text document, out of which the keyphrases were extracted, is linked to the topic map.
- *Information retrieval on topic map:* While the first two use cases deal with the extension of the knowledge base (*knowledge base input*), this use case describes the functionality to search upon the topic map in order to find media content (*knowledge base output*).

2. The actor "Domain expert" is generally the expert of a special knowledge domain (e.g. politics, business or sports) in the media company:
   - *Create & Manage topic map:* The domain expert is responsible for the creation and change management of the knowledge base comprising the maintenance, restructuring and enhancement of the topic map. In regular meetings, the editorial staff and domain experts discuss potential adjustments to the structure of the topic map.

In the following depiction of the system architecture of the fully-implemented prototype KeyTEx, we will focus on the first three use cases concerning the system's interfaces to the actor "Editor".

KeyTEx can be characterized by the classical three-layer architecture shown in Figure 4. The components of the different layers can be integrated in one local standalone application or deployed in a distributed client-server environment based on a thick-client approach. The three-layer architecture will serve as a guiding principle to structure the rest of this section.

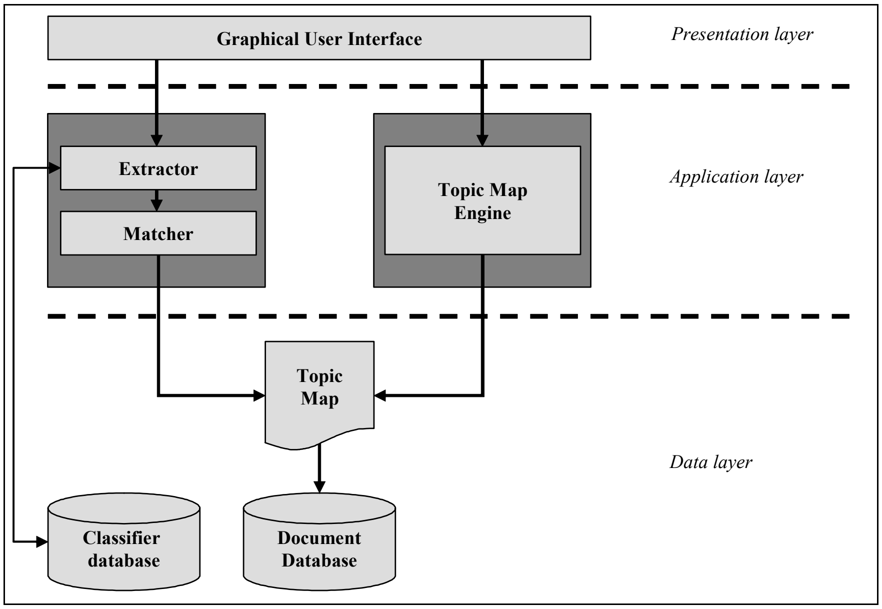
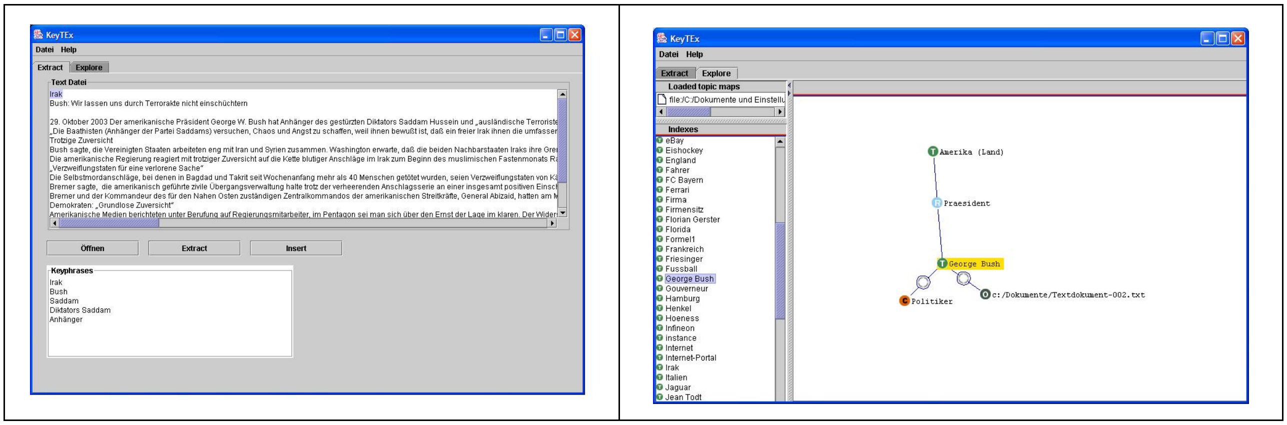


Figure 4. KeyTEx system architecture

At the top of the architecture, an integrated graphical user interface (*presentation layer*) provides consistent access to the functionalities of the prototype. The GUI for the extraction functionality enables the user to integrate text documents, to trigger the extraction process, to adjust manually suggested keyphrases and to insert them into the underlying topic map (see Figure 5 left). The GUI on top of the topic map presents topics as nodes, associations as edges and text documents as occurrences (see Figure 5 right). If the user clicks on a node, the direct neighbor topics and associated occurrences with unified resource information are shown.



Irak
Bush: Wir lassen uns durch Terrorakte nicht einschüchtern

29. Oktober 2003 Der amerikanische Präsident George W. Bush hat Anhänger des gestürzten Diktators Saddam Hussein und „ausländische Terroriste
„Die Baathisten (Anhänger der Partei Saddams) versuchen, Chaos und Angst zu schaffen, weil ihnen bewußt ist, daß ein freier Irak ihnen die umfasser
Trotzige Zuversicht
Bush sagte, die Vereinigten Staaten arbeiteten eng mit Iran und Syrien zusammen. Washington erwarte, daß die beiden Nachbarstaaten Iraks ihre Grei
Die amerikanische Regierung reagiert mit trotziger Zuversicht auf die Kette blutiger Anschläge im Irak zum Beginn des muslimischen Fastenmonats Ra
„Verzweiflungstaten für eine verlorene Sache"
Die Selbstmordanschläge, bei denen in Bagdad und Takrit seit Wochenanfang mehr als 40 Menschen getötet wurden, seien Verzweiflungstaten von Kä
Bremer sagte, die amerikanisch geführte zivile Übergangsverwaltung halte trotz der verheerenden Anschlagsserie an einer insgesamt positiven Einscl
Bremer und der Kommandeur des für den Nahen Osten zuständigen Zentralkommandos der amerikanischen Streitkräfte, General Abizaid, hatten am M
Demokraten: „Grundlose Zuversicht"
Amerikanische Medien berichteten unter Berufung auf Regierungsmitarbeiter, im Pentagon sei man sich über den Ernst der Lage im klaren. Der Wider

Irak
Bush
Saddam
Diktators Saddam
Anhänger

Figure 5. Keyphrase extraction GUI (left) and topic map GUI (right)

In the middle (*application layer)*, the logic of the functionalities is bundled into three different components:

- *Extractor:* The keyphrase extraction component is based on the open-source WEKA libraries[2] and is implemented in Java. The Extractor scans the text document and resorts to a trained classifier database in order to decide upon the relevance of each scanned keyphrase for selection.
- *Matcher:* The main purpose of the Matcher is to provide the connection between the keyphrase extraction and topic map visualization functionalities. As soon as the extracted keyphrases of a mined text document are submitted by the editor, the Matcher compares each keyphrase with existing topics of the topic map. If a keyphrase matches with a topic, the URI (Unified Resource Identifier) of the text document is added to the occurrence tag of the identified topic. If there is no match, the keyphrase is written into a log file for later consideration. A UML activity diagram summarizes the functionality of the Matcher (see Figure 6).

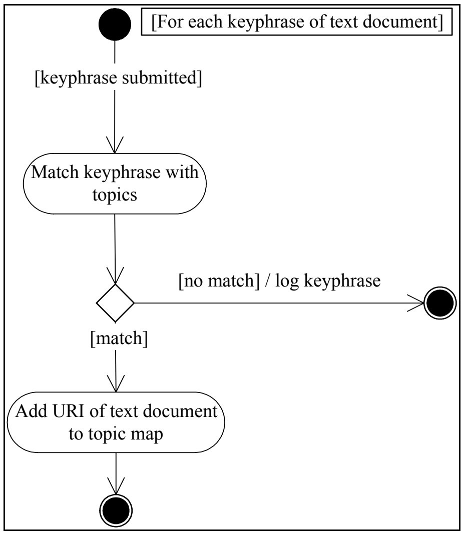

[For each keyphrase of text document]
[keyphrase submitted]
Match keyphrase with topics
[no match] / log keyphrase
[match]
Add URI of text document to topic map

Figure 6. Functionality of the Matcher component

- *Topic Map Engine:* The topic map engine of KeyTEx adopted the open-source toolkit tm4j[3] to realize the visualization of and interactive information retrieval on the underlying XML-based topic map. It is built upon an event-based architecture that provides topic map processing and browsing functionalities.

[2] Developed at the University of Waikato, New Zealand: http://www.cs.waikato.ac.nz/~ml/weka/index.html

[3] www.tm4j.org

The *data layer* at the bottom of the architecture consists of three components. The *classifier database* that comprises a hash table with keyphrases and associated probability values, enables the Extractor to determine and suggest relevant keyphrases for the application domain. The hash table itself is the result of a continuous training process with a set of domain-specific documents based on machine-learning algorithms. In our work, the probability-based Naïve Bayes rule was chosen because of its adequacy for text categorization (Joachim, 1998). The *document database* stores the physical text documents referenced by occurrence tags in the topic map. The *topic map*, as the central ontology of the application domain, represents the connecting base between the Extractor/Matcher and Topic Map Engine (see Figure 4). An extract of the topic map showing the three basic elements (topics, associations and occurrences) is illustrated in Figure 7.

```
<?xml version="1.0"?>
<topicMap xml:base="http://www.wi.bwl.uni-muenchen.de/default.xtm"
xmlns:xlink="http://www.w3.org/1999/xlink">
   ...
  <topic id="Foreign Minister">
    <instanceOf>
      <topicRef xlink:href="#Politician"/>
    </instanceOf>
    <baseName>
      <baseNameString>Foreign Minister</baseNameString>
    </baseName>
  </topic>
   ...
  <topic id="Colin Powell">
    <instanceOf>
      <topicRef xlink:href="#Foreign Minister"/>
    </instanceOf>
    <occurrence id="id013">
      <resourceRef xlink:href="c:/Dokumente/Textdokument-004.txt">
      </resourceRef>
    </occurrence>
    <baseName>
      <baseNameString>Colin Powell</baseNameString>
    </baseName>
  </topic>
   ...
  <association id="isMinister">
    <member>
      <roleSpec>
        <topicRef xlink:href="Foreign Minister"/>
      </roleSpec>
      <topicRef xlink:href="Colin Powell"/>
    </member>
    <member>
      <roleSpec>
        <topicRef xlink:href="Country"/>
      </roleSpec>
      <topicRef xlink:href="USA"/>
    </member>
  </association>
</topicMap>
```

Figure 7. Topic map extract

LAB EXPERIMENT FINDINGS

In order to assess the economic usefulness of the *information retrieval* functionality of KeyTEx, a lab experiment was conducted in January 2004[4]. The main thesis of the experiment was that the browsing performance in network-based

[4] The text mining functionality will be empirically investigated in an upcoming lab experiment.

information topologies is higher than in hierarchical ones depending on the search setting (see chapter 1)[5]. Since we have developed the KeyTEx prototype as a supporting tool for a general population consisting of editors and archivists in media companies, a sample of randomly selected students with a major in Information systems and New Media Management seemed representative enough to generate preliminary results. Half of the sample had to browse in the classical Windows NT Explorer[6] (control group), the other half in KeyTEx (experimental group) on a semantically comparable ontology and document base. In statistical terms, the information topology (NT Explorer: hierarchical tree structure; KeyTEx: network-based XML topic map) represented the central independent variable influencing the browsing efficiency of the users measured as search performance (search time, number of clicks) within given search tasks. The four types of search questions posed to the experimental and control group (see Table 1) were developed on the basis of the integrated model of browsing and searching of Choo et al. and several times adjusted in pretests (Choo, Detlor and Turnbull, 2000). The logic behind the sequence of the search questions is to increase the browsing complexity. On the one hand a growing number of search items within one search task had to be found, on the other hand the semantic level of abstraction of the search items was raised.

Search target(s)		**Search question**
1.	One search item, high level of specificity	"Find information about Michael Schumacher"
2.	One search item, low level of specificity	"Find information about Internet criminality"
3.	Three search items, high level of specificity	"Find information about the Iraq war and corresponding statements of the American and German Foreign Minister"
4.	Three search items, low level of specificity	"Find information about top managers in business, politics and sports"

Table 1. Search target(s) and corresponding search questions

With one *specific* search item in their mind, test users navigated faster to the search target in Windows NT Explorer than in KeyTEx (search task 1). The opposite could be noticed for search task 2 where KeyTEx users found the search target in almost a third of the time compared to Windows NT Explorer test users. Search tasks 3 and 4 produced results that are to some extent contradictory to the results of search task 1 and 2. KeyTEx test users fulfilled search task 3 ("Set of interrelated, *specific* search items within *similar* topic contexts") significantly faster than Windows NT Explorer test users, whereas the opposite occurred in search task 4 ("Set of interrelated, *unspecific* search items within *different* topic contexts"). Interestingly, an *statistically significant* difference between the means of the search performance measures of the two groups could be discovered in search tasks 2 and 3 in which KeyTEx users proved to browse more efficiently. Detailed results of the lab experiment are illustrated in Table 2.

The mental assignment of the specific search item "Michael Schumacher" to a folder name proved to be easier in the NT Explorer hierarchy than in the semantic network of KeyTEx (search task 1). This result can possibly be attributed to the better orientation in well-known and familiar hierarchical structures or to the difficulties test users had with finding a good enough entry network node. In search task 2, the level of abstraction of the search item was raised, because the term 'Internet criminality' was less tangible and could intuitively be associated with different overarching topics. This increased browsing complexity irritated NT Explorer users showing feelings of confusion while clicking up and down the hierarchical information structure. KeyTEx users, however, clicked rapidly through different semantic network paths to the search target following their first intuitions. Specifically in this search scenario, the saying "Many roads lead to Rome" bears analogy to semantic networks.

	Group 1	*Group 2*	*t-test parameters*	
	n=10	n=10	t-value	p
	KeyTEx	***Windows NT Explorer***		(*p<0,05)
1. Search target: One search item, relatively high level of specificity				

[5] The underlying research question was explicitly not to compare the *quality of retrieved information items*, but solely the *information retrieval speed* in order to gain insight into the usefulness of network-based information retrieval in specific search scenarios.

[6] The parametric, keyword-based search functionality in the Windows NT Explorer was deactivated.

Mean search time (in sec.)	28,0	23,0	-1,849	0,081
Mean number of clicks to fulfill search task	4,0	3,4	2,714	0,014*
2. Search target: One search item, relatively low level of specificity				
Mean search time (in sec.)	38,0	95,4	3,043	0,007*
Mean number of clicks to fulfill search task	3,0	15,2	2,955	0,008*
3. Search targets: Three search items, relatively high level of specificity				
Mean search time (in sec.)	64,2	91,5	1,932	0,069
Mean number of clicks to fulfill search task	8,7	12,9	4,541	0,000*
4. Search targets: Three search items, relatively low level of specificity				
Mean search time (in sec.)	76,8	50,3	-1,477	0,157
Mean number of clicks to fulfill search task	8,3	10,5	1,075	0,296

Table 2. Experiment results: Mean search time and mean number of clicks with KeyTEx and NT Explorer

In search tasks 3 and 4, the complexity was increased once again by integrating a greater number of search items into one search task. Although the level of indeterminateness of the search items grows from search task 3 to 4, browsing in the network-based information topology didn't prove to be more efficient than browsing in the hierarchical information topology. A possible reason for these results could be that with a greater number of search items in *different* topic contexts, the distance between the starting and ending network node, one has to cover, increases. In addition to that, a general handicap of networks is that there is no consistent organizing principle for the semantic links between network nodes. Whereas relationships between parent and children folders in hierarchical information structures can consistently be described as taxonomic (is-kind-of/is-a) and meronymic (part-of) hierarchies (Cruse, 2000), semantic links between equal network nodes can represent various inconsistent relationships (e.g. "is-a-relationship" or "has-a-relationship", etc.). Since users are not given consistent hints about which links lead to *far-off* network nodes, they often end up with a "lost in space"-feeling.

The above results don't allow a final interpretation of the advantages and disadvantages of network-based browsing compared to hierarchy-based browsing. However, we could demonstrate that there are search settings where browsing in network-based information topologies is at least as efficient as browsing in hierarchical information topologies. As preliminary results, we propose that those search settings can be described with the following characteristic features:

- *Single search item with a high degree of indeterminateness ("brainstorming search setting")*: Because of the indeterminateness of the task, the navigator feels uncertain about the way to the search goal. As networks support natural associative thinking, users can do some brainstorming while navigating through the semantic network. Especially when editors or archivists are searching for background material (e.g. for a feature story or a documentary), browsing on semantic knowledge networks could be conducive to creative thinking and serendipitous discovery.
- *Multiple (interrelated) search items with a low level of specificity ("hopping or hub-and-spoke search setting")*: Associative browsing in network-based information topologies prove to be more efficient when users are searching for a chain of documents about semantically proximate topics (hopping from one topic to another related topic within the same category) or about one main topic with several peripheral topics (hub-and-spoke searching). These advantages are partly inherent in the structure of networks due to the high likelihood of semantically similar node neighbors, partly due to the better fit with the conceptual model of the navigating user. This kind of search settings can for instance be found in media companies when editors or archivists are searching for archival material concerning recurrent big media events (e.g. The Annual Academy Award as main topic with movies and celebrities as peripheral topics).

CONCLUSION AND POTENTIAL FOR DEVELOPMENT

With regard to the prototype KeyTEx, we have presented an architecture that integrates both keyphrase extraction and network-based information retrieval techniques on the basis of XML topic maps. Through the development of a functional prototype that was grounded on a conceptual input-output-model, we have shown that the combination of both functionalities is technically feasible. In order to provide insights into the economic usefulness and applicability of the prototype's information retrieval component, we presented first experimental findings on the advantages of semantic-based information retrieval (with KeyTEx) over hierarchy-based information retrieval in specific search settings. Although the results were not entirely unambiguous and stringent, they point to time-saving potentials in search settings with one unspecific search item and several interrelated search items within similar topic contexts. These time-saving potentials due to enhanced IT capabilities can eventually be conducive to process and quality improvements in content management workflows in (media) companies (e.g. Santhanam and Hartono, 2003).

Although we have provided first empirical findings regarding topic map based information retrieval, further empirical tests have to be carried out concerning both functionalities of the prototype. Regarding the presented experimental findings, further experiments must be conducted with varying search questions, various underlying semantic ontologies and additionally controlled variables (e.g. intelligence or previous knowledge of users) in order to enhance internal and external validity. As far as both functionalities of KeyTEx are concerned, the application of the whole prototype in different real-life systems (e.g. search engines, content management systems, etc.) and application scenarios (e.g. digital asset management, e-learning, knowledge management) with a larger sample size should be investigated, too. As a next central step in our action research project, we will apply KeyTEx in a publishing house.

For the future much work remains to be done. We want to highlight two major issues. First, the keyphrase extraction functionality could be extended from pure text mining to content-based indexing of pictures, films and audio-files. Although picture, video and audio mining technologies have not yet achieved results as marketable as text mining technologies, further progress can be expected in these research areas. Another potential for development can be seen in the usage of topic map query engines (e.g. TMQL). A query engine would improve the usability of KeyTEx since users would not only be able to browse in the topic map but also to perform parametric queries to the system. These two possible improvements are only a drop in the ocean, but indicate that more research in the intersection of (semi-)automated content-based indexing and network-based information retrieval is needed.

REFERENCES

1. Adams, K. (2002) The Semantic Web, *Online,* 26, 4, pp. 20-23.
2. Baeza-Yates, R. and Ribeiro-Neto, B. (1999) *Modern Information Retrieval,* ACM Press, New York.
3. Batra, S., Bishu, R. R. and Donohue, B. (1993) Effects of Hypertext Topology on Navigation Performance, *Proceedings of the Fifth International Conference on Human-Computer Interaction, Elsevier*, pp. 175-180.
4. Biezunski, M., Bryan, M. and Newcomb, S. R. (1999) ISO/IEC FCD 13250:1999-Topic Maps. ISO/IEC JTC 1/SC34.
5. Böhm, K., Heyer, G., Quasthoff, U. and Wolff, C. (2002) Topic map generation using text mining, *Journal of Universal Computer Science,* 8, 6, pp. 623-633.
6. Booch, G., Rumbaugh, J. and Jacobson, I. (1999) *The Unified Modeling Language User Guide,* Addison-Wesley Professional, Reading, Massachusetts.
7. Chen, H., Lynch, K. J., Basu, K. and Ng, T. (1993) Generating, Integrating, and Activating Thesauri for Concept-Based Document Retrieval, *IEEE Expert, Special Series on Artificial Intelligence in Text-Based Information Systems,* 8, 2, pp. 25-34.
8. Choo, C. W., Detlor, B. and Turnbull, D. (2000) Information Seeking on the Web: An Integrated Model of Browsing and Searching, *First Monday,* 5, 2.
9. Cruse, D. A. (2000) *Meaning in Language. An introduction to semantics and pragmatics,* Oxford University Press, New York.
10. Daum, B. and Merten, U. (2002) *System Architecture with XML,* Morgan Kaufmann.
11. Ester, M. and Sander, J. (2000) *Knowledge Discovery in Databases,* Springer, Heidelberg.
12. Feldman, R. and Dagan, I. (1995) Knowledge Discovery in Textual databases (KDT), *Proceedings of the 1st international conference on knowledge discovery, Montreal*, pp. 112-117.
13. Fluit, C., Sabou, M. and van Harmelen, F. (2002) Ontology-based Information Visualisation, In: *Visualising the Semantic Web* (Ed, Geroimenko, V.) Springer Verlag.

14. Hearst, M. A. (1997) Text data mining: Issues, techniques and the relationship to information access, *Presentation Notes for UW/MS workshop on data mining*.
15. Hearst, M. A. (1999) Untangling text data mining, *Proceedings of the 37th Meeting of the Association for Computational Linguistics, College Park, MD*, pp. 3-10.
16. Hearst, M. A., Elliot, A., English, J., Sinha, R., Swearingen, K. and Yee, K. (2002) Finding the Flow in the Website Search, *Communications of the ACM,* 45, 9, pp. 42-49.
17. Hovy, E. and Lin, C. Y. (1999) Automated Text Summarization in SUMMARIST, In: *Advances in Automatic Text Summarization* (Eds, Mani, I. and Maybury, M.) MIT Press, Cambridge, pp. 81-94.
18. Joachim, T. (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features, *Proceedings of ECML-98, 10th European Conference on Machine Learning, Chemnitz, Germany,* Springer, pp. 137-142.
19. Kohonen, T. (2001) *Self-Organizing maps,* Springer, Berlin et al.
20. Lee, J.-H., Kim, Y.-G. and Yu, S.-H. (2001) Stage Model for Knowledge Management, *Proceedings of the 34th Hawaii International Conference on System Sciences*, pp. 7071.
21. Maedche, A. and Staab, S. (2000) Semi-Automatic Engineering of Ontologies from Text, *Proceedings of the 12th International Conference on Software Engineering and Knowledge Engineering(SEKE2000), Chicago, IL*, pp. 231-239.
22. Markus, M. L., Majchrzak, A. and Gasser, L. (2002) A design theory for systems that support emergent knowledge processes, *MIS Quartely,* 26, 3, pp. 179-212.
23. Mohageg, M. F. (1992) The influence of hypertext linking structures on the efficiency of informational retrieval, *Human Factors,* 34, pp. 351-367.
24. Pepper, S. (2000) The TAO of Topic Maps: finding the way in the age of infoglut, Last Visited: 02/02/2004, http://www.gca.org/papers/xmleurope2000/papers/s11-01.html
25. Santhanam, R. and Hartono, E. (2003) Issues in linking information technology capability to firm performance, *MIS Quartely,* 27, 1, pp. 125-153.
26. Sebastiani, F. (2002) Machine Learning in Automated Text Categorization, *ACM Computing Surveys,* 34, 1, pp. 1-47.
27. Shneiderman, B. (1998) *Designing the user interface: strategies for effective human-computer interaction,* Addison Wesley, Reading, MA.
28. Witten, I. H. and Eibe, F. (1999) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations,* Morgan Kaufmann.