Automatically Building Concept Structures and Displaying Concept Trails for the Use in Brainstorming Sessions and Content Management Systems

Christian Biemann, Karsten Böhm, Gerhard Heyer, and Ronny Melz

University of Leipzig, Institute of Computer Science
{biem,boehm,heyer,rmelz}@informatik.uni-leipzig.de

Abstract. The automated creation and the visualization of concept structures become more important as the number of relevant information continues to grow dramatically. Especially information and knowledge intensive tasks are relying heavily on accessing the relevant information or knowledge at the right time. Moreover the capturing of relevant facts and good ideas should be focused on as early as possible in the knowledge creation process.

In this paper we introduce a technology to support knowledge structuring processes already at the time of their creation by building up concept structures in real time. Our focus was set on the design of a minimal invasive system, which ideally requires no human interaction and thus gives the maximum freedom to the participants of a knowledge creation or exchange processes. The initial prototype concentrates on the capturing of spoken language to support meetings of human experts, but can be easily adapted for the use in Internet communities that have to rely on knowledge exchange using electronic communication channels.

1 Introduction

With a growing number of communities in the Internet, tools are needed that provide aid in communication within and between them. People with different backgrounds might use the same term for different concepts or call one concept with different names – often without even noticing. The result is an inherent misunderstanding between people who want and need to cooperate, leading to communication problems, frustrations and finally to financial losses.

Our goal is to support the communication and mutual understanding in two ways: On one hand we provide a visualisation tool for spoken language, having its application in informal creative and usually highly innovative meetings like brainstorming sessions or open space workshops (cf. [Owen, 1998]. In these scenarios the tool does not only serve as an automatic documentation method by arranging the keywords in a meaningful way, but also provides corpus-based associations in order to enrich the conversation with concepts that are related but might have been forgotten by the participants.

On the other hand we use the same visualization engine for displaying single documents as trails on a so-called *semantic map*. The idea of a semantic map is heavily relying on the well known concept of geographical maps, in which visual structuring of interesting locations and the emphasizing of relevant paths between them are the key concepts to provide an orientation for the user. Other important properties of geographical maps that deal with an appropriate information filtering are the different

T. Böhme et al. (Eds.): IICS 2004, LNCS 3473, pp. 157–167, 2005.

levels of scale and thematic scopes (e.g. political maps vs. hiking maps). The semantic maps introduces in this paper will reflect this properties too. Since the location of the concepts on the map is fixed, users can grasp the contents of a document rapidly and compare documents in a visual way.

The description of these features and the discussion of possible applications will be the topic of this paper. Both features are part of a prototype implementation called “SemanticTalk” which has been presented to the public in at the worlds largest IT-Exhibition CeBit in Hanover, Germany, in spring 2004.

The remaining paper is organized as follows: In the following section we will introduce the underlying technologies for our approach. Section 3 describes the current implementations in the prototype. Current applications are shown in section 4. The concluding section 5 shows some issues for further research.

2 Base Technologies

After introducing the notion of statistically significant co-occurrences, we describe how the results of this calculation method – the global contexts – give rise to the construction of semantic maps and the automatic associations provided by the system in brainstorming mode.

2.1 Statistically Significant Co-occurrences

The occurrence of two or more words within a well-defined unit of information (sentence, document) is called a co-occurrence. For the selection of meaningful and significant collocations, an adequate co-occurrence measure has to be defined: Our significance measure is based on a function comparable to the well-known statistical *G-Test* for Poisson distributions: Given two words *A*, *B*, each occurring *a*, *b* times in sentences, and *k* times together, we calculate the significance *sig(A, B)* of their occurrence in a sentence as follows:

$$\begin{aligned} & sig(A,B) = x - k\log x + \log k! \\ & \text{with } n = \text{number of sentences}, \\ & x = \frac{ab}{n}. \end{aligned}$$

Two different types of co-occurrences are generated: based on occurrence *within the same sentence* as well as *immediate left and right neighbors of each word*. For further discussion on co-occurrences, see [Biemann et al. 2004].

In short, the co-occurrence statistics result in connection strengths between words, which tend to appear in the same contexts. In the following, connections below a significance threshold are dropped, leading to the notion of the global context of words: Whereas words occurring with a reference word in a single sentence are called local contexts, the global context of the reference consists of the most significant co-occurrences.

2.2 Visualization with TouchGraph

TouchGraph (see also *www.touchgraph.com*) is a freely available 2D visualization tool for dynamically developing complex graphs. It is written in Java and provided under the conditions of the apache-style software[1]. Within our solution the software is used as a framework for the visualization of semantic maps.

The layout mechanisms of TouchGraph provide methods to add, change/delete both nodes and edges to a graph already being visualized, causing it to jiggle and find a new stable state. The principle can be captured if you consider the nodes to be positive electrical charges, dispersed due to the electrical force but some of them bounded by a constraint force imposed by the edges. This is exactly the effect we want to visualize: unbounded objects (which can be whole clusters or single word forms) are drifting away until they find a stable position, because semantically they do not have anything in common. Different nodes representing semantic objects can be connected with one or more semantic relations while the network grows. As a result the distance between them decreases, nodes related to each other are automatically repositioned. Aside from the automatic layout mechanism, TouchGraph provides generic zoom- and other scrollbars and a context menu for user interface that can be used to parameterize the shape of the displayed network.

The following enhancements were added to the basic functionality, mostly motivated by our application of visualizing word semantics:

- various zoom scrollbars (see section 2.3)
- colored nodes and edges depending on semantic categories (see section 3.1)
- a possibility to disable the convergence towards total equilibrium, thus “freezing” the graph in its current layout

Furthermore, the necessity arose to add new edges without changing the positioning of the graph (see section 3.2), which could be obtained by freezing it.

As the main panel only represents a detail of the whole graph, a function dynamically scrolling and zooming has been implemented to keep track of the red thread.

2.3 From Co-occurrences to Semantic Maps

Based on methods described in [Faulstich et al. 2002], it is possible to extract keywords that reflect important concepts of a document collection describing a domain automatically and language-independently. For this analysis, based on differences in relative corpus frequency, a large reference corpus, such as the Wortschatz Corpora, containing the result of an analysis of a large, representative document collection in German language (see http://www.wortschatz.uni-leipzig.de) is needed.

These keywords serve as a basis for generating a semantic map of the domain: by analyzing the co-occurrence of words within the same sentence statistically (see [Quasthoff et al. 2003]), and techniques of visualization (see [Schmidt 2000]), it is possible to display the keywords on a two-dimensional plane in a way that relatedness between keywords is reflected by short distances on the map.

By use of this method the keywords arrange themselves in clusters. All keywords of one cluster are related to a single event. Keywords can be part of several clusters, but are painted only once during the visualization process. As a consequence they will be located between the clusters they belong to. Figure 1 shows a cluster for “Tastatur” (keyboard), and the two clusters of “Monitor” (monitor), being part of the cluster related to computer hardware and a TV magazine called “Monitor”.

[1] The software can be obtained at *http://prdownloads.sourceforge.net/touchgraph*.

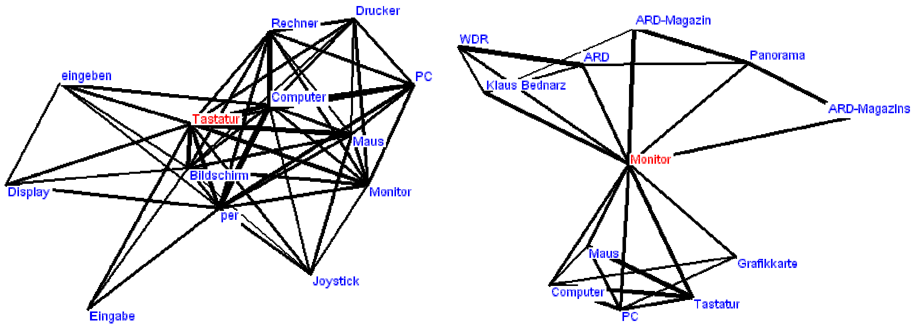

Fig. 1. One cluster for “Tastatur” (keyboard) and two clusters for “Monitor” (monitor)

Note, that the graphs in the figure above are centered on the reference words (“Tastatur”, “Monitor”). This causes the effect that the cluster from the left graph in figure 1 is being reproduced in the right graph of the figure in a somewhat distorted way. As opposed to this, when visualizing the semantic map of the whole domain, the relative position between the keywords remains the same.

The automated creation of the semantic map for a specific domain consists of the following processing steps:

1. Extracting keywords using a relative corpus frequency comparison
2. Adding global context sets from the domain
3. Adding connections between words that are members of each other’s global context

The size of the resulting map can be parameterized by the number of extracted keywords and the size of the global context sets.

Naturally, the number of important keywords is varying with the size of the text collection that has been used to define a domain. Especially when thinking of large domains, like newspaper archives or document collections of companies, a single screen for displaying the whole map will clearly not suffice for showing all the keywords in lexicalized form. As a practical solution, we introduced two windows for visualizing the same semantic map at two different zoom factors: the “topic survey window” serves as an overview of the whole domain, the “local context window” displays highly granular relations between specific content units. For a screenshot, see figure 2.

To adjust the granularity of the display according to the user’s needs, three rulers have been implemented:

- *Conceptional zoom:* display of nodes as dots (see topic survey window) vs. lexicalized display i.e. as words in the local context window, larger words are keywords with higher rankings
- *Granularity:* the total number of nodes displayed
- *Scale:* ratio of the size of the local context compared to the whole semantic map

The visualization of the semantic map is pre-calculated for a specific domain and serves as constant background knowledge to visualize the content-path of each spoken sentence in the semantic net. When extending the domain, the map has to be recalculated. An initialization of the visualization software with positional information of the previous domain map ensures resemblance to the previous map on order to facilitate re-orientation: When e.g. visualizing newspaper content, politics will stay in the lower right corner while sports is always in the upper middle part.

3 SemanticTalk – A Tool for Innovation Acceleration

As mentioned in the introduction, the SemanticTalk tool supports two operation modes: a *brainstorming mode* in which a semantic map will be constructed dynamically from the input and a *red-thread mode* that visualizes a communication trail in a pre-calculated map. Both modes will be described in detail in this section. The input for the system can be obtained from different sources: it may be loaded from text files, directly typed in using a provided input field or spoken text recorded from a headset. For the conversion of spoken text, we use *Linguatec*'s VoicePro 10[2], which is speaker-dependent general-purpose dictation software for German language.

The SemanticTalk user-interface that represents the different views of the semantic map as well as the controls that modify the behaviour of the tool is show in the figure below.

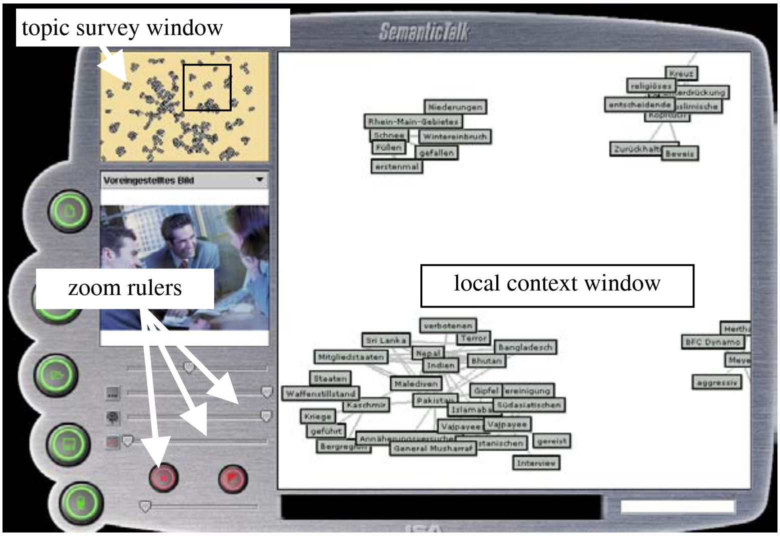

Fig. 2. The *SemanticTalk* user interface. The local context window is a zoomed variant of the topic survey window and can be changed with the three zoom rulers. Other controls: New, Load, Save, Exit, Microphone on/off, Freeze graph, Brainstorming / Red thread Switch, Input text bar and scrolling

3.1 Visualization of Associations in Brainstorming Mode

Project planning and realization usually starts and is further accompanied by brainstorming sessions where several people simply talk about possibilities of what to do.

[2] For more information, visit: http://www.linguatec.de

In the beginning, every participant has an isolated view on the issue, whereas the goal of the brainstorming session is the exchange of the views in order to share the same view afterwards. In brainstorming mode, the software acts as a common brain that contains the views of all participants and proactively gives associations itself.

For preparation, a text collection of the domain has to be processed as described in [Biemann et al. 2004] to provide the terminology and the global contexts to the system. If existent, ontology or some typological system can be loaded in order to assign types to concepts and relations.

The brainstorming session starts with a blank screen. Words of the conversation that are considered to be important (an easy strategy is to use only nouns of a substantial frequency in the domain corpus) are thrown into the visualization. These words are in the following referred to as *core words*. If two words stand in the global context relation they are connected, which leads to an arrangement of concepts that reflects semantic relatedness by smaller distance. Each word is painted in a single frame with white background, indicating the human source.

The global contexts of words serve as candidate sets for *associations*. These global contexts are over-generating, first of all because the sets are usually too large, second because of the lexical ambiguity of words that leads to mixed semantics in the global contexts: i.e. when speaking about bank transactions, one would not want a system to associate river (bank) or park-bench related words. The strategy for associations is as follows: only words that appear in two ore more global contexts of core words are displayed and connected to their related core words.

When choosing associations, preference is given to typed relations, assuming that relations from the ontology are of higher interest to the brainstormers than other, untyped relations. The number of associations as well as the number of necessary core words for displaying an association can be parameterized, making the system tuneable with respect to associative productivity. The frames of associated words are painted in grey background colour, indicating machine source.

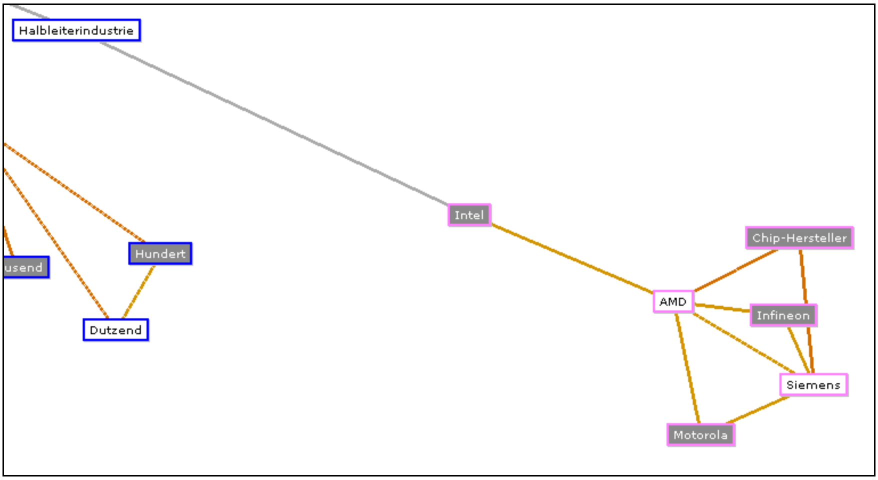

Fig. 3. Brainstorming Mode: core words and associations

Figure 3 shows an example for some semiconductor companies: *AMD, Halbleiterindustrie* (semiconductor industry) and *Siemens* are core words and *Motorola, Infineon, Intel* and Chip-Hersteller (chip manufacturer) have been associated by the system. All company names are coloured in pink, reflecting organisations, the typed connections between the companies reflect co-hyponomy, between *Chip-Hersteller* and two companies the IS-A-relation holds. The connection between *Halbleiterindustrie* and *Intel* is not typed. Typization of words and relations can either be loaded from an existing domain-specific ontology or annotated manually or semi-automatically.

The resulting graph can be exported via a template engine into XML-like description languages for example RDF to be used and further refined in domain specific applications. Depending on the target structures an export into other graph and concept-oriented knowledge structures, such as Topic-Maps or ontologies are possible.

An important aspect of the export functions will be the domain- or role specific filtering, using typed nodes an edges which can be used to build contextualized knowledge structures. Possible applications are the knowledge management applications that rely on a contextual information provision that could be gained from a business process, for example (see [Hoof et al. 2003]).

3.2 Visualization of Concept Trails on Semantic Maps in Red Thread Mode

Semantic maps (see section 2.3) provide a good overview of a domain by visualizing keywords and their relatedness. While they are extracted from a document collection and serve as model for the whole domain, it is possible to display single documents as paths through them.

We visualize the content of the broadcast as trajectory in the pre-calculated semantic map. Words of interest from the document are marked in red and are connected in sequence of their occurrence by directed edges. Here we distinguish between two kinds of connections (see figure 4):

1. A connection between two concepts has already been present in the semantic map: the two keywords are semantically related and therefore connected by a red colored line.
2. The connection has not been present in the semantic map: this indicates a shift of topic, just like the introduction of a new piece of information. These kinds of connections are drawn in orange.

When using spoken input, the colored connections are visualized in real time; the local context window is moving in a way that the most recent keyword always is located in the middle of the plane. For text file input, the resulting train is pre-calculated and visualized after processing the whole text file.

Words that are contained in the semantic map are marked in red in the input. The input document gets connected to the semantic map, which gives rise to bidirectional retrieval: the semantic map's local context window adjusts to (clickable) red words in the input, and document contexts from the single document and from the underlying collection can be retrieved by selecting a word from the semantic map.

By using a fixed semantic map and dynamically representing input as indicated, it is possible to examine

- coverage of domain: the more clusters are visited, the more the document covers the domain

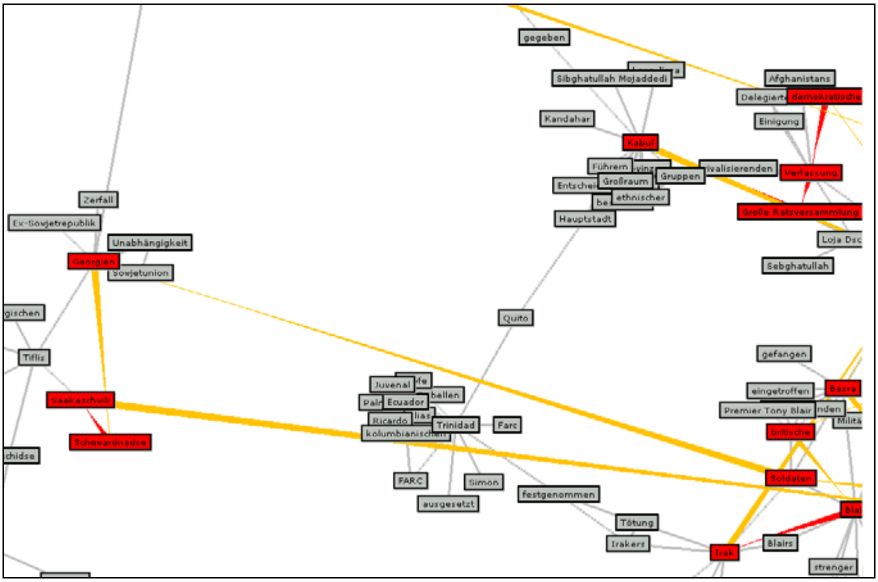

Fig. 4. Red thread visualization: topic shift to and from Georgian politics (left part of the window). Other clusters visited: Afghanistan (right upper corner) and Iraq politics (right lower corner), not visited by red thread: Trinidad and Columbia politics (lower middle). Data basis: *Tagesschau* broadcast news in red on one-day newspaper corpus in grey

- coverage within topics: the more words in a cluster are marked in red, the more extensively the document treats the topic of the cluster
- relatedness of document to domain: few red words in the input indicate non-relatedness
- contiguity of document: many successive long range orange connections indicate semantic incoherence
- comparison of different documents: either in a visual way by displaying several documents in different colors or by using semantic map concepts as features for classification or clustering

The great advantage of this representation lies in displaying what is described in the document, and the same time, what is *not* being dealt with. By using a fixed representation for the domain, the practiced user can grasp the document's content in a few seconds and then decide whether to read it or not.

4 Applications

To some extent this model captures important aspects of our cognitive processing of language. One natural application will therefore be in the area of natural language dialogue systems, in particular in Question-Answering systems: When evaluating a path in a semantic net for certain key concepts alternative continuations to the path actually followed become apparent. This indicates possible questions that the system might ask back to the user. We might also evaluate the actual path with respect to its information value (with reference to the domain specific conceptual expectations), and better structure the dialogue accordingly.

A further application may be found in the area of content and knowledge management. By mapping a document onto the semantic background net (instead of spoken natural language input), we get an indication of the topics that it treats, and of the topics that it does not cover as well. This may be used to evaluate, for example, the coverage of technical descriptions (e.g. instructions for the use of technical devices/user manuals). Also, by evaluating the conceptual path, documents may be compared with respect to the way they verbalize conceptual relations. Thus the SemanticTalk can be seen as a tool supporting the digestion of electronic documents, which in turn will help to find the relevant information within a large document collection faster and more easily.

With respect to Internet Communities we are currently evaluating another promising area: the analysis of high volume discussion forums in its various forms (Mailing lists, newsgroups, Wikis and blogging systems). Many of these community-centric systems propose a significant entry barrier to a newcomer in the community, since the sheer amount of information and its dynamic change makes it difficult to find the relevant information or follow the communications threads within an established community. In order to apply the SemanticTalk we will carry out a background analysis of a complete high volume Mailing list (to be used as a semantic map) and use single positing or communication threads as inputs for the red thread functionality to highlight the mentioned topics and concepts and draw relevant association from the background knowledge.

The full paper will present some result of the experiments currently carried out with some high volume mailing lists and newsgroups.

5 Related Research

The general issues of supporting group meetings with IT-systems are dealt with in [Krcmar et al. 2001] at a general level, whereas the extraction of theme structures from spoken language relates to different research areas and can be seen as the task of information or topic extraction (see [Cowie & Lehnert 1996]) and [Grishman 1997]). Information extraction from spoken language was discussed in [Palmer et al. 1999] but focuses on the template-based extraction of named entities. Another interesting approach in which the computer plays the role as a mediator within a session is illustrated in [Jebara et al. 2000], although the scope of this approach is limited to a restricted set of trained topics using a machine learning approach. The proposed solution is not limited to a set of topics and therefore only dependent on the analysis of an appropriate document collection in the background.

To our knowledge there is no related work of visualizing concepts in semantic maps automatically. While [Bergmann & Dachs 2003] visualize documents on maps due to relatedness of extracted keywords, the content description of the documents is inserted manually and rather high conceptual. Other approaches for visualization of semantics, like [Mertins et al. 2003] or [Paier 2003] who both display organisational structures, assume a well-defined database or ontology, whereas this work finds its data basis by statistical analysis.

6 Future Work

For the future, we plan to extend the tool in different ways. We identified another necessary zoom method: A simple hierarchical clustering algorithm could group the concepts into sensible clusters using their closeness on the map. When using the cluster-zoom, several close concepts are collapsed into a multi-node with a circular label. This multi-node is positioned at the center of gravitation to preserve the topology of the map. Lexical labels for multi-nodes can be chosen by heuristics from the labels of the collapsed nodes; when in red thread mode, the grade of redness indicates the fraction of red nodes in a multi-node. Multi-nodes should be expandable and collapsible by automatic means, but also by the user who defines a personalized view on a fixed, underlying map.

For improving usability, we will implement a variety of import formats, in as well as provide means to export the resulting graphs into machine-readable markup-files (e.g. in RDF Format) and printable graphics formats.

References

1. Bergmann, J. and Dachs, B. (2003): Mapping Innovation in Services. A Bibliometric Analysis. Proceedings of I-KNOW'03 International Conference on Knowledge Management, Graz, Austria
2. Biemann, Chr.; Bordag, S.; Heyer, G.; Quasthoff, U.; Wolff, Chr. (2004): Language-independent Methods for Compiling Monolingual Lexical Data, Proceedings of CicLING 2004, Seoul, Korea and Springer LNCS 2945, pp. 215-228, Springer Verlag Berlin Heidelberg
3. Cowie, J.; Lehnert, W. (1996): Information Extraction. In: Communications of the ACM. Vol. 39, Nr. 1, S. 80-90.
4. Faulstich, L.; Quasthoff, U.; Schmidt, F.; Wolff, Chr. (2002): Concept Extractor - Ein flexibler und domänenspezifischer Web Service zur Beschlagwortung von Texten, ISI 2002
5. Grishman, R. (1997): Information Extraction: Techniques and Challenges. In: Maria Teresa Pazienza (Hrsg.): Information Extraction, Lecture Notes in Artificial Intelligence. Rom: Springer-Verlag.
6. van Hoof, A.; Fillies, C.; Härtwig, J.: Aufgaben- und rollengerechte Informationsversorgung durch vorgebaute Informationsräume, Aus: Fähnrich, Klaus-Peter; Herre, Heinrich (Hrsg.): Content- und Wissensmanagement. Beiträge auf den LIT'03. Leipzig (2003).
7. Jebara, T.; Ivanov, Y; Rahimi, A.; Pentland, A. (2000): Tracking conversational context for machine mediation of human discourse. In: Dautenhahn, K. (Hrsg.): AAAI Fall 2000 Symposium - Socially Intelligent Agents - The Human in the Loop. Massachusetts: AAAI Press.
8. Krcmar, H., Böhmann, T., & Klein, A. (2001). Sitzungsunterstützungssysteme. In G. Schwabe, N.A. Streitz, & R. Unland (Eds.), CSCW Kompendium - Lehr- und Handbuch für das computerunterstützte kooperative Arbeiten. Heidelberg: Springer.
9. Mertins, K., Heisig, P. and Alwert, K. (2003): Process-oriented Knowledge Structuring. Proceedings of I-KNOW'03 International Conference on Knowledge Management, Graz, Austria and Journal of Universal Computer Science (JUCS), Volume 9, Number 6, Pp. 542-551, Juni 2003
10. Owen, H. (1998): Open Space Technology: A User's Guide, 1998, Berrett-Koehler Publishers Inc., San Francisco

11. OpenspaceWorld: Portal on OpenSpace information in the Internet: http://www.openspaceworld.org
12. Paier, D. (2003): Network Analisys: A tool for Analysing and Monitoring Knowledge Processes, Proceedings of I-KNOW'03 International Conference on Knowledge Management, Graz, Austria
13. Quasthoff, U., Richter, M., Wolff, C. (2003): Medienanalyse und Visualisierung: Auswertung von Online-Pressetexten durch Text Mining, in Uta Seewald-Heeg (Hrsg.), Sprachtechnologie für die multilinguale Kommunikation, Beiträge der GLDV-Frühjahrstagung 2003, Gardez!-Verlag, Sankt Augustin
14. Schmidt, F. (2000): Automatische Ermittlung semantischer Zusammenhänge lexikalischer Einheiten und deren graphische Darstellung, Diplomarbeit, Universität Leipzig