Concept Indexing

Angi Voss, Keiichi Nakata, Marcus Juhnke
GMD-FIT
Schloss Birlinghoven
D-53754 Sankt Augustin, Germany
<first name>.<surname>@gmd.de

ABSTRACT
Marking text in a document is a convenient way of identifying bits of knowledge that are relevant for the reader, a colleague or a larger group. Based on such markings, networks of concepts with hyperlinks to their occurrences in a collection of documents can be developed. On the Internet, marked documents can easily be shared, concepts can be constructed collaboratively and the concept-document network can be used for navigation and direct access. Text marking, grounded concepts and the Internet as base technology are characteristics of our tool for managing so called "concept indexes". We describe the current and the new design and outline some application scenarios: electronic help desks, information digests on the Web, teaching design in virtual classes and planning under quality control in distributed teams.

Keywords
Knowledge management, documents, concepts, collaboration, software agents, text marking

MOTIVATION

Marking Text in Documents

When people read an important document, they often highlight, underline or annotate key passages. This is a very convenient technique and can be applied equally well while quickly skimming through some pages or studying them carefully.

The markings may serve several purposes: to quickly recover the key issues when browsing through the paper later on, to memorize key terms and phrases, or to pass one's comments together with the document to colleagues or friends. The marked pieces may be targets of bookmarks, of index entries, or they may be cut and pasted into a glossary, a group thesaurus or onto cards in a stack. Such knowledge organizers may be created for oneself, e.g. while preparing a book, they may be created for a customer, e.g. by an information analyst, or they may be shared within a group, e.g. a project team.

GROUP 99 Phoenix Arizona USA

We believe that the benefits of marking text in a document can be increased further when the documents are not on paper but in electronic form, especially when the Internet is used as a medium. At the same time, the individual efforts for text marking remain comparatively low.

- First of all, the Internet is a vast source of electronic documents: from Web Sites, email and news archives, digital libraries or shared workspaces. Search engines, lists of links and Web catalogs simplify the search for interesting documents.
- Indexes, glossaries and more recent knowledge organizers, like concept maps, idea maps or knowledge maps, may be hyperlinked to targets in the documents, yielding a navigable knowledge space, a Web of knowledge.
- On the Internet such a hyperspace can easily be shared with others, it can even be constructed collaboratively in distributed groups. Circulation of (marked) electronic documents to interested parties is much easier and faster than circulation of paper.
- The documents on the Web can automatically be monitored for changes, triggering updates to the whole structure in order to keep it consistent. The updated content can be presented on dynamically generated Web pages.
- Improvements in OCR allow interpreting markings from paper, and hand-held scanning appliances can be used on paper for electronic marking.

A Web-Based Tool

In the following we will present our tool for communicating knowledge based on markings of text in documents on the Web. The tool treats the markings as indicators of bits of information or knowledge that the readers perceive in the documents. Marked text pieces are taken as occurrences of concepts, and each concept is automatically linked to all occurrences of similar text in other documents. Thus, each concept provides a list of cross-references, and the entire set of concepts serves as an index to the collection of documents. Users may further

structure and reorganize the concepts, thereby creating a space for knowledge exploration. Each document in this space is dynamically augmented on display by highlighted and hyperlinked concept occurrences.

We call the structures created with the help of our tool "concept indexes" rather than "knowledge spaces", because knowledge is actually only in the heads of people (or internal to artificial agents). However, documents are a means of communicating knowledge: writing a document is a form of knowledge externalization and reading a document is a form of knowledge internalization [17]. For a community of readers a concept index is a means to indicate and communicate the knowledge they perceive in a collection of documents.

Sharing and Collaboration

The target users of our tool are persons with a common interest or task, teleworkers, mobile workers and distributed teams who would profit from making use of the knowledge possessed by each other. We expect that their number will rapidly increase with continued globalization, decentralization, and flattening of communication structures.

Knowledge is context-dependent. Whether one perceives some data as information or knowledge depends on his or her purposes, interests, and experiences. The value of communicating some data as "handles to knowledge" depends on both, the producer and the recipient. Often people accept their individual interpretations and conclusions only partially. An approach where one person reads the documents and derives a recommendation for all others may only work in well-established groups. In very loose co-operations, people may simply share their document collections and leave it to each other to read the documents and form their own opinions. In the middle of the spectrum there are communities where an individual would share views with selected other persons only.

Therefore, our tool will capture the relation between contributions and contributors. It will provide configurable social filters as instruments to choose information, reconstruct remote contexts, appreciate others' contributions, and identify experts. This will promote social orientation and facilitate trust in team building and collaborative action.

CONCEPT INDEXES

Features

A concept index is accessible to a set of registered users. It maintains of a collection of documents and a set of interrelated concepts. Concepts are cross-referenced with explicitly marked pieces of text and with implicitly detected similar pieces of text. Two major windows, shown in figures 1 and 2, offer different options for manipulating a concept index. A separate frame is devoted to each option (the enumeration below corresponds to the numbering of the frames in figures 1 and 2):

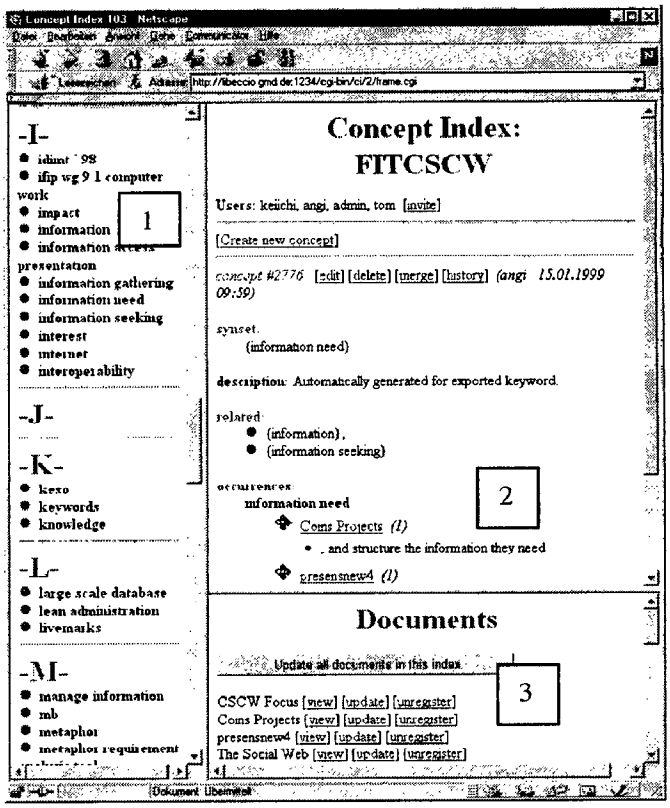

Fig.1. A concept index with the list of terms (textual expressions) (left frame "1"), the list of documents (bottom right frame "3"), and the properties of a concept (top right frame "2").

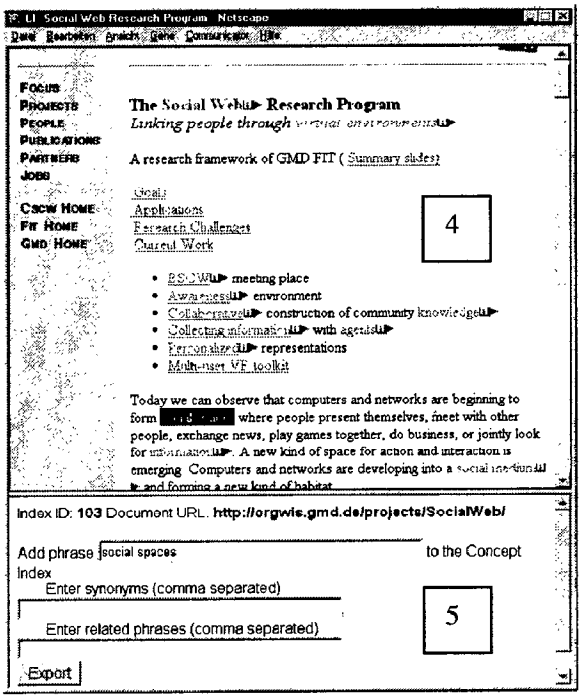

Fig.2. A document with "colored" occurrences of text pieces assigned to concepts and icons for links to concept definitions (top frame "4"). To add new textual expressions (terms etc.) to concepts, text pieces can be highlighted using the mouse, or the form (bottom frame "5") can be used.

1. For orientation, one can use a listing of the key terms that have been marked. Each marking leads to the description of the associated concept (in frame 2).
2. On can focus on a particular concept and obtain a listing of key text pieces that relate to this concept. They lead to the respective source document (in frame 4).
3. One can inspect the document collection and contribute new documents. A document is presented as in shown in frame 4.
4. When a document is inspected, text markings and cross-references are dynamically inserted so that the presentation reflects the view of the community.
5. One can mark pieces of text in the document, associate them to concepts and create relations between the concepts. (Another window allows editing of concepts in detail.)

Nature of the Concepts

In a concept index, any object, conceptualization, idea, person, place etc. can be a concept. A concept can textually be expressed in terms of words, phrases, sentences, paragraphs etc., and also by other concepts.

Concepts are Textually Grounded

A concept may have a number of textual expressions, which may differ in their grammatical construction, terminology or jargon, language, and content. In our tool, assigning a piece of text to a concept is a simple action that triggers a text analysis and a subsequent extensive search for similar text pieces in the documents. A flexible matching is achieved through stemming and lemmatization, stop-word recognition, and word order permutation within a match window of an appropriate size.

Users can influence the manner in which normalization is performed by specifying the way textual expressions are interpreted. In the current design, we allow users to specify phrases (consisting of multiple words, the sequence of which is significant), proper names (not to be lemmatized) and concepts. A user may specify a part of the textual expression as a concept to emphasize that it is not the exact words that are important, but the concept they express, or more simply, the meaning is important. This has the effect of replacing the text with alternative texts that express the same concept.

Concepts are Related to Each Other

Concepts may be related in many ways. We presently support simply two relations, and their distinction is rather practical: users may wish to group several concepts and form a new concept out of them; this is a case of "comprise" relation. When two concepts are simply seen to be closely associated, but not necessarily to be grouped into another concept, such a relation can be captured as an "associated" relation.

Concepts as Knowledge Handles

In a distributed, possibly heterogeneous community of information seekers and providers we cannot expect a common terminology and should tolerate independent contributions from individuals to the concept structure as far as possible. One consequence is that we keep the number of relations small and impose a weak semantics on the relations. That is, "comprise" and "associated" relations, while there exist informal definitions, only capture the intuitive relations between concepts. If certain properties like cycles in the relations are not desired, tools could be provided to detect them and initiate a process for consensus among the users.

As another consequence, concepts are not defined, but hermeneutically grounded [24]. Readers must interpret a concept in the context of their occurrence and usage in the documents. The context, i.e., the set of actual instances of a concept appearing in a document collection (called *occurrences*) is automatically extended through the search for similar pieces of text in the documents. As texts in different jargons or languages can be associated to the concepts, they may serve as a cross-community and cross-language thesaurus.

DESIGN AND IMPLEMENTATION

Early in 1998, our team was constituted as part of a research program called "The Social Web"[1]. This program explores the possibilities of transforming the Internet from an information space into a social space where people appear from behind the information and form virtual communities. Concept indexes are our team's contribution to the collaborative construction of knowledge in the context of a Social Web.

Our tool profits from several recent technological advances:

- The Internet, the World-Wide Web and associated languages and protocols;
- multilingual electronic thesauri and text preprocessing tools for machine translation;
- data mining, text mining, and information extraction techniques to automatically obtain information and knowledge from text;
- software agents that operate in the background as distributed and asynchronous software processes;
- groupware for creating social awareness through history summaries, synchronous and asynchronous notifications, and communication and interaction between users.

The tool is implemented as a client-server architecture on the Internet. It is developed in Java with HTTP communications between clients and a server primarily

[1] http://orgwis.gmd.de/projects/SocialWeb

through Java Servlets. Concepts, documents and user data are stored in a relational database. To cater for the changes in documents, only document URLs are stored.

Access to Indexes

The access to the concept index server is controlled by supplying the user name and password. A user can create an index, and can share it with other users by inviting them to the index, hence the access to an index is restricted to its members. Each index has its own concepts and document collections.

Creating/updating a New Concept

When a user registers a text piece, e.g., by highlighting in a browser, and if it does not currently exist in the working index, a new concept is created in the index. A concept has a set of synonyms (a *synset*) which "describes" the concept, and links to related concepts ("comprise" and "associated" relations), and its occurrences in the document collection. Upon the creation of a new concept or addition of a new text piece in the synset of an existing concept, the occurrences of text pieces are detected by accessing the URL of documents in the collection.

Occurrence Detection

The occurrence detection procedure ensures that 1) for each word, its lemmatized form matches that in the document, and 2) for a multiple-word text piece (e.g. a phrase) each word, except stopwords, appears within a predefined proximity in any order, again matching the lemmatized form. While this carries the danger of detecting occurrences other than desirable ones, we observed that the benefit of detecting occurrences in many possible ways outweighs the nuisance of these undesirable matches. This occurrence detection is performed by software agents. The agents are implemented in Java on our agent platform SOaP [25] and operate asynchronously on possibly distributed machines.

Modifying/editing a Concept

In the current version, the slots in the concept descriptions listed above can be modified or updated using an HTML form. Using this form, users can specify concept relations, add new text pieces to the synset, etc., and all such modifications and editing activities are recorded in a log. The log can later be used to trace the changes in the index, and retrieve information such as who added new concepts and relations, and when.

Document Display

Since only the URL of the documents are stored, each time a document is opened for display, it is processed to insert HTML tags to display concept occurrences in a distinctive color (using the same concept occurrence detection procedure described above) and a clickable icon that provides hyperlink from each of the detected concepts to its concept description page.

The current prototype of our concept index management tool, available since fall 1998, is the first robust, operational prototype [18]. It has been used to explore several application scenarios that will be sketched below. In the current version, concepts cannot be named or comprise other concepts and can only meaningfully be assigned to small segments of texts like phrases. Feedback from the first users gave rise to a new design in spring 1999, which includes extended agent services and more groupware features. The implementation of this version is currently under way.

Agent Services

In spite of the weak semantics, a lot of work is required to keep the index up-to-date, enrich it and achieve a high quality information management. Some of this work is extensive but routine, and ideally suited for delegation to software agents, which operate asynchronously and persistently in the background. The tasks agents perform in our tool are as follows.

Document Monitoring

A concept index is not a shared workspace; it does not store documents, but records URLs and file locations and retrieves them as needed (trading storage for asynchronous processing time). Agents will monitor the documents at their external sources and, when they detect changes, trigger updates of the concept occurrences. Thus a concept index can be kept up-to-date with any changes in the documents.

Concept Occurrence Detection

Concept occurrences have to be updated when the documents change, when documents are registered and unregistered to the document collection, when concepts are created, and when the textual expressions assigned to a concept are changed. Periodically or triggered by these events, agents will fetch the documents and analyze them for concept occurrences.

Document Retrieval

The textual expressions assigned to a concept can be processed to provide search expressions for similar text pieces in the document collection. These search expressions can also be passed to agents that retrieve new documents from external sources.

For this purpose we will reuse agents from LIVEMARKS [25], our system of agents for collaborative information retrieval which is also based on the SOaP platform. LIVEMARKS services have already been integrated with shred workspaces in BSCW [2].

Document Assessment

The documents in a concept index as well as newly retrieved documents can be analyzed for different purposes. For example, the relevance of a document can be estimated due to the number of concept occurrences in the document, and its relevance can be guessed by the type of document, author, source etc.

Concept and Relation Detection

By statistically analyzing the documents based on the existing concepts, new terms and key phrases, concept clusters and associated concepts can be detected. Agents can suggest them to the users who may choose to ignore, accept or reject them. Since such suggestions are typically made by a user, this is a case in which agents behave like one of the members of the user group.

Groupware Features

To increase social awareness, and to tailor a concept index to a particular task, user or subgroup, our tool will

- provide various kinds of event summaries that allow to relate contributions to the concept index to their contributors, and thus better judge their social relevance for the reader;
- allow commenting on concepts and documents in order to exchange assessments in the community;
- provide powerful and composable filtering and sorting criteria, and means to structure the document collections and the whole index of concepts. Concept indexes tend to grow quickly and subindexes are the recommended means of tailoring the knowledge to one's particular interest. Subindexes allow focussing on particular themes, document subsets, "what's new", "what's most popular", or "what's most authoritative".
- In addition, we plan to offer advanced awareness services through the use of NESSIE, an application-independent awareness server under development in our research group [21]. A NESSIE server may receive event descriptions from arbitrary sources on the Internet, and it offers generic synchronous and asynchronous event indicators, like ticker tapes, or summary emails. Indicators can be configured for particular types of events and be integrated into Web pages, like static home pages or dynamically generated pages for a concept index.

DISTRIBUTED CONSTRUCTION OF CONCEPT INDEXES

A concept index imposes no particular construction procedure. Documents, concepts, and relations can be added, modified or deleted in any order.

Nonetheless, we can distinguish elementary steps, corresponding to elementary roles in the distributed construction of a concept index, and a default workflow for starting the construction. These steps are as follows.

1. Collection of documents.

A number of relevant or typical documents are identified and possibly organized into subcollections. This seed set may later be extended manually, by agents, or by some other push process. Persons who contribute documents play the role of "information providers".

2. Identification of key text.

While browsing and reading the documents key pieces of text are identified. The process is comparable to highlighting a document on paper. Currently, we are considering several mobile interaction devices such as reading pens that scan lines of text paper, or display tablets that operate with pens and either include a computer or are connected to one (see fig. 3). In a basic computer screen interface, highlighting can be simulated with the mouse in a document browser.

3. Formation of concepts

If the target concept for a new piece of text is not explicitly designated, a new concept will be created by default. With the number of text pieces, the number of concepts will grow prompting the need for imposing some structure. Concepts can then be merged and relations be introduced. Agents can make helpful suggestions, for instance, by analyzing concept occurrences in documents for co-occurring concepts and suggesting associations between them. Persons that shape the concept network play the role of "information organizers".

4. Sharing viewpoints

A concept index usually grows fast. Document subcollections and comprising concepts create shared structures. Subindexes, which are defined by a subset of documents and concepts, introduce sharable viewpoints. They may focus, for example, on a particular topic, on approved information, on "what's new" or "what's changed". The creation of subindexes is another organizing activity. Selectors and sorters, which users may dynamically apply, are means to tailor the information presentation to their personal preferences.

5. Reviewing

Users can comment on concepts and documents in order to exchange and discuss opinions. To underpin the discussion, the usage of a concept index can be logged and evaluated to determine the popularity of concepts, documents, and subindexes. Users who assess a concept index play the role of "reviewers".

6. Management

Concepts and documents may be introduced tentatively and later, on the basis of comments and log data, be discarded or officially accepted. This is the job of a "editor", who is also responsible for overall coordination.

Fig. 3. Candidate appliances for mobile interaction with concept indexes: reading pens (*Quicktionary* from WizCom Technologies Ltd., above) and book size computers with display tablets and pens (*Stylistic 2300* from Fujitsu Personal Systems Inc., below). (Pictures obtained from manufacturers' Web site and reproduced with their permission.)

APPLICATION SCENARIOS

The aim of the tool is to provide a functionality that anyone can contribute in any way. Nonetheless, different application scenarios may favor a particular combination of roles and a distribution of roles to distinguished kinds of users. As a consequence, we expect a need for tailored user interfaces for the different applications.

Information Digests

Increasingly, the Web is offering valuable information: techniques and methodologies, pilot projects and practical experiences, funding organizations, social and legal matters, issues of competition, regional, national and international activities. As a first access point to obtain such information, citizens, decision makers, consultants, and researchers prefer to consult the Web sites of appropriate trusted organizations.

One example are the Clearing Houses which the Convention on Biological Diversity (Rio de Janeiro, 1992), requires to be established in each of its over 160 member states. In spite of a pilot phase from 96-98, initiated at the third member states' conference in Buenos Aires, today's Clearing Houses still fall short of the requirement to offer "demand-driven, up-to-date information on biodiversity that is of high quality, and to present it in a user-oriented way, at neutral costs".

These requirements are neither met by a virtual library, which is collected and indexed by librarians, nor by an automatically collected full-text indexed archive, nor by standardized meta-keywords which authors may optionally attach to their pages. For instance, guidance for information seeking users is only available through pre-specified categories, whose inflexibility may undermine efficient information exploration and user satisfaction. However, the Web is a global information space, its content is produced and consumed by heterogeneous groups with diverse interests and jargons that are too dynamic to be anticipated by any standardization.

Concept indexes address these problems and help the construction of a "digest" of high quality information. Without imposing a fixed terminology a priori, grounded concepts can be created as needed and thus capture an emerging terminology in the community. Grounded concepts are easy to create from the users' free-text queries or by experts and editors while they read a document.

An information digest in the form of a concept index can evolve through collaboration between editors, experts and information seeking users. Involvement of the users guarantees that the digest meets their needs, while the contributions of experts and their peer review guarantees a high quality. This distribution of work, according to the competencies and interests of the participants substantially reduces individual efforts. In particular, the work of the editor is reduced so that the costs of maintaining an information digest are affordable.

This scenario has been outlined together with ZADI, the operator of the German Clearing House for biodiversity.

Planning Under Quality Control

During complex planning processes a multitude of requirements, constraints, forms, guidelines, regulations, norm, catalogs, and precedents have to be considered.

In order to verify the quality of a planning process, it is necessary to document which of these inputs have actually been taken into account at what occasions. If the planning process itself is documented in terms of electronic discussions and output documents, these could be linked to their input documents.

In order to improve the quality of a planning process, it should be possible to request related documents and also to automatically obtain suggestions for potentially relevant documents.

A concept index can give this support: Planners can easily express a request by highlighting a critical piece of text when they are writing a paper or note, and they automatically obtain suggestions through the text detection and cross-referencing mechanism.

Together with an international chemical enterprise, we intend to explore this scenario for the development of the first plan of a chemical process control system. Such a plan

involves decisions of high impact that may have to be aware of a large number of other, often text-based documents. The planning process will be conducted through electronic discussions, the output documents will be communicated through shared workspaces (using ZENO [9]), and the referenced documents will be available from repositories on the Intranet and from catalogs and norms on CD.

Help Desk

At an electronic help desk, users and consultants exchange emails to solve a problem. Usually, the emails are archived for future reference and reuse. Recurring questions may be extracted and edited for a FAQ. An online manual may be a third source of advice. The email archive, FAQ and manual represent different kinds of abstractions, and a piece of advice may move upward in the hierarchy as it is recognized to be of general interest.

Help desks are typically operated in shifts, and by consultants with different expertise. Assignment to a help-desk may be considered as a kind of training-on-the-job for new colleagues. If there is not much capacity for a help desk, the email archive can be made public and users may play the role of volunteer consultants, according to their expertise.

In any case, a person acting as a consultant may not be aware of all information in the archive, FAQ or manual. To better benefit from the stored advice, access must be supported in a pragmatic way, i.e., it should be easy to associate a query in an email with those pieces of text in the documents which represent appropriate answers.

For this purpose, the experts must provide knowledge in the following form: First of all, symptomatic phrases must be identified in the queries; second, these symptoms must be grouped according to the class of problems, which they indicate. Problem classes must be associated to solution classes, which in turn must be connected to pieces of text that describe solutions.

In the field of case-based reasoning, [13] has tried to support such text-based problem solving. They stress that knowledge acquisition is a considerable factor.

In a concept index, the knowledge can be expressed in terms of concepts for symptoms, problems, and solutions. Emails, FAQ and online manual constitute the document collection. When an email arrives at the help desk, all occurrences of symptoms are automatically highlighted and cross-referenced with occurrences of solutions to the associated problems. The cross-references may automatically be sorted by relevance. The consultants may browse this list and select or edit the best replies. The new answer is automatically indexed and archived.

If there is no appropriate reply, the email may be raising a new problem. It must be passed to an expert, who will identify the key phrases and assign them to a symptom. If a suitable symptom does not exist, a new one must be created and associated with a corresponding problem concept. Further, if such a problem concept does not exist, it must be created and associated with a solution. If necessary, a new solution must be introduced and associated with suitable pieces of texts in the documents. If there are no appropriate text pieces, a completely new reply must be formulated to answer the question.

We hope that this knowledge extension process can be done on the fly, with little additional effort while the experts answer a query for two reasons: (1) the concepts in a concept index can be created as needed, that is, as new problems are identified, and (2) the concepts are not formally defined, but need only be exemplified through their usage in the documents.

This application scenario is the result of our discussions with the help desk team of the BSCW shared workspace system.

Teaching Design

Design has been classified into routine, innovative and creative tasks [4]. Creative solutions introduce new solution spaces, innovative solutions extend existing solutions spaces, and even for routine tasks, the space of solutions may be tremendously huge. Therefore, design is often done by adapting former cases to the new requirements. Teaching design is similarly done using previous cases, especially for architecture and engineering [1], [20]. There are several attempts to formalize cases and support design by retrieving or even adapting similar cases for a new problem [16], [8].

However, formalizing a design case is a complex task, and confronts a principal problem: It is never clear what aspects of a design may become relevant in the future. Therefore, a design may not appropriately be indexed and hence ignored for a new problem.

A concept index addresses both problems. A design case need not be formalized, but is described by a set of documents and CAD plans that would be produced as ordinary teaching material. Documents may contain requirements, refer to regulations, describe the design history or rationalize CAD plans, and CAD plans may focus on different aspects, parts, or versions. Tutors then introduce the concepts that they intend to teach using the cases. Concepts may be associated with pieces of text, which will automatically be detected in the documents, or they may manually be associated with pieces of CAD plans. Students may explore such a concept index in order to solve a given task.

Furthermore, students may be asked to document their design on a pad. On the pad, they drag, drop and connect the concepts they consider relevant, even if they lead to dead-ends or remain unexplored variants. The activity on the pad may be monitored by automated design assistants. At any time, such an agent may try to propose a list of new concepts that are best associated with the concepts on the

pad. For that purpose, they use the "associated" relation between concepts, which should be stronger when two concepts occur close to each other in many documents.

We are discussing this application scenario in a tele-tutoring context that involves design tutors from different architecture and engineering schools. The tutors would first create concept indexes for their own cases and the concepts they would consider important. Then they would exchange or merge their concept indexes and apply the others' concepts to their own cases and vice versa. On this basis, they can try to compare their concepts and discuss a common framework. They will also see their own cases under new perspectives, namely indexed by formerly unanticipated concepts. Even students could introduce new concepts that would re-index (and re-interpret) the existing cases. This exploits the fact that documents and indexes in a concept index may be introduced in any order and circumvents the chicken-and-egg problem between cases and concepts.

RELATED WORK

As the range of applications suggest, concept indexes combine aspects of various tools for knowledge and information management.

Like an ontology, a concept index has concepts and relations, c.f. [11] or the ONTOLOOM[2] editor. But since relations and concepts are more informal, a concept index is rather a proto-ontology, or precursor of an ontology. This aspect has been raised by the developers of the ONTOBROKER tool [7].

Compared to a thesaurus like WORDNET [6], the number of different relations in a concept index is low. Instead, it ties the concepts to pieces of texts in the documents and supports structured concepts. Thus, a concept index could serve as a precursor for a more elaborate group-specific or task-specific thesaurus.

The cross-references between pieces of text allow the use of a concept index like a word index or cross-reference list, and enhance it by a conceptual organization and support for larger text segments. For example, the Dynasites[3] tool automatically creates links from words used in one document to a central glossary, and from the glossary back to where the word was used in other documents.

Concept maps or knowledge maps, as supported by tools like THEBRAIN[4] or MINDMANAGER[5], graphically connect concept nodes and are applied to structure ideas, to access background material attached to the nodes, and to communicate this knowledge. Concept indexes can serve the same purposes: while they lack the graphic interface, they actively detect connections between concepts and text pieces in the background material.

With tools for qualitative research, like ATLAS.TI[6] or WINMAX[7], users can select pieces of text, assign them to a category, and query for particular pieces of texts. But the tools do not detect similar occurrences of text pieces or, based on the distribution of occurrences, suggest new relations or concepts.

Tools such as IBM's text analysis tools[8], Oracle's CONTEXT CARTRIDGE[9] or Semio's text mining tool[10] do extensive text analysis in order to extract knowledge, but there is a danger of producing junk. In a concept index, the primary knowledge sources are the users and automatic knowledge extraction will be applied only to suggest extensions to that knowledge.

Tools for semantic retrieval expand queries and filters results [19], [3], [10] and [14] generate queries from text and link the results to the documents. We combine both approaches, and additionally analyze the new documents at the text level and integrate them into the existing concept structure.

Most tools mentioned above support collaboration through the sharing of knowledge. Social filtering systems [22] move from simple knowledge sharing to recommendation. Examples are GROUPLENS for filtering news articles [12], COMMENTOR for filtering shared annotations [23], and JASPER for filtering bookmarks [5]. COMEMO generates virtual conversations by combining associations contributed by different people [15].

To summarize, a concept index has a unique combination of features, which individually may be found in other kinds of tools. Our choice of features has been tailored to contextualized and conceptualized knowledge which we expect to be valuable for re-use in many ways.

DISCUSSION

In this section we discuss some issues which can be identified in the current approach of concept indexes, some of which can be addressed in future work.

Dealing with Obsolete Data

In the current design, new concept entries and relations can be added manually by users or through assistance using text mining tools. While there is a danger that tools such as text

---

[2] http://www.isi.edu/isd/ontosaurus.html

[3] http//Seed.cs.colorado.edu/

[4] http://concepts.thebrain.com/

[5] http://www.MindMan.com/english/product

[6] http://www.atlasti.de

[7] http://www.winmax.de

[8] http://www.software.ibm.com/data/iminer/fortext/tatools.html

[9] http://www.oracle.com/st/o8collateral/html/xctx5bwp.html

[10] http://www.semio.com/mining.html

mining tools produce undesirable entries, users can be selective in accepting suggestions from these systems, and might benefit from letting the system perform statistical analyses of documents to enrich the index. Hence the argument over the applicability of such tools is, in essence, a trade-off problem.

The more difficult issue in maintaining a concept index is perhaps that of deleting entries. As in the case of groupware, users may be reluctant to delete entries created by other users, or simply put less effort in deleting obsolete data compared to adding new data. This is not a simple problem to deal with, and offering automation or system support would be a challenge.

Capturing Dynamic Changes of Interests

Since a concept index is expected to evolve through contributions from the users sharing it, it would also reflect the way interests among the users change over time. From the record of events that took place during the development of an index, such as addition and deletion of concept entries, concept relations and documents, we can reconstruct the history of drifts of interests, if any, within the group.

Such a service is not only useful for tracking user interests and offer information to create incentives to form new groups and subgroups, but also potentially applicable to the analysis of document-based collaboration. We can associate concepts with the documents that motivated their introduction to the index, and identify influential documents in that group. Services of this sort involving the analysis and interpretation of the event log will be examined in the future.

Maintaining User Motivation

The success of a groupware like the concept index management tool depends on how it can raise and maintain motivation at the individual level. At the moment, without an empirical basis, we can only speculate on this issue. We expect that there may be different solutions for the different applications. In general, it will be crucial that every user personally benefits from the tool. Apart from social prestige, the satisfaction of personal knowledge needs would be such a value. Knowledge needs may be satisfied faster by a concept index than by a stand-alone tool, because everyone may profit from the others' contributions. Additionally, a concept index accumulates social information; it constructs a community view that a stand-alone tool could never offer.

SUMMARY

To summarize, a concept index may satisfy a range of information and knowledge needs:

What is the content of the document collection?
Users can introduce concepts and relations between them as a means for content-oriented navigation through the documents.

What do the concepts represent?
Concepts are not formally defined, but instead assigned pieces of text that express them in the documents. All possible occurrences of concepts are dynamically spotted by matching similar text pieces in the document collection. Thus, each concept is indexed to its meaning implicit in its usage in the documents.

What are the nuggets in the documents?
Text pieces in a document that are assigned to the concepts can be seen as a source of shared knowledge. Document presentation can expose such nuggets by highlighting concept occurrences.

Are there hidden connections between the documents?
Documents are implicitly related to each other by common concepts. Text pieces that belong to common or related concepts are automatically cross-referenced so that users can easily browse among them.

What are relevant documents?
Documents can be registered to a document collection by the users or retrieved by the system, which generates queries from the concepts. The number of concept occurrences can indicate the relevance of retrieved documents.

ACKNOWLEDGEMENTS

We would like to thank several persons who helped to develop application scenarios for concept indexes: Horst Freiberg and Frank Begemann at ZADI (http://www.zadi.de), Wolfgang Ahrens and Martin Garre from BAYER AG (http://www.bayer.de), Alberto Giretti from University of Ancona (http://idau.unian.it/idau/idau.htm), Rudolf Ruland and Thomas Koch from OrbiTeam (http://www.orbiteam.de), Tom Gordon from Dialogis (http:www.dialogis.de). Volker Paulsen and Thomas Kreifelts from our research group are helping to integrate the concept index tool with LiveMarks and SoaP. We extend our thanks to anonymous reviewers whose comments contributed to the improvement of the paper.

REFERENCES[11]

1 Alexander, Ch., Ishikawa, S. Silverstein, M. *A pattern language: towns, buildings, construction*, New York, NY, Oxford University Press, 1977.

2 Appelt, W., Hinrichs, E. and Woetzel, G., Effectiveness and efficiency: the need for tailorable user interfaces on the Web, *WWW7 Conference*, 1998, http://www7.scu.edu.au/programme/fullpapers/1830/com1830.htm827-846.

3 Borgo S., Guarino N., Masolo C. and Vetere, G. Using a Large Linguistic Ontology for Internet-Based Retrieval

[11] Links verified January 1999

of Object-Oriented Components, in *Proc. of 9th Int. Conf. on Software Engineering and Knowledge Engineering* SEKE 97, (Madrid, Spain, 1997), http://www.ladseb.pd.cnr.it/infor/Ontology/ontology.html

4 Chandrasekaran, B. Towards a Functional Architecture for Intelligence Based on Generic Information Processing Tasks, *Proceedings of IJCAI 87*, 1987, 1183-1188.

5 Davies, J., Weeks, R. and Revett, M. JASPER: Communicating information agents for the WWW, in *Proc. 4th Int. World Wide Web Conf.* (Boston MA, Dec. 1995), *World Wide Web Journal 1*, 1, O'Reilly, Sebastopol CA, 1995, 473-482.

6 Fellbaum C. *An Electronic Lexical Database*, MIT Press, 1998, http://www.cogsci.princeton.edu/~wn/

7 Fensel, D., Erdmann, M. and Studer, R., Ontobroker: The Very High Idea, *Proceedings of the 11th International Flairs, Conference* (FLAIRS-98, Sanibal Island, FL, May 1998), http://www.aifb.uni-karlsruhe.de /WBS/broker/

8 Gebhardt, F., Voss, A., Graether, W., Schmidt-Belz, B. *Reasoning with complex cases*, Kluwer Academic Publishers, 1997.

9 Gordon, T., Karacapilidis, N. and Voss, H. ZENO - a mediation system for spatial planning, in U. Busbach, D. Kerr, and K. Sikkel, (eds.), *CSCW and the Web - Proceedings of the 5th ERCIM/W4G Workshop* (Sankt Augustin, Germany, February 7-9, 1996), GMD Technical Reports No. 984, 55-61, http://nathan.gmd.de /projects/zeno/zenoSystem.html

10 Green, S. Automated link generation: can we do better than term repetition? in *7th International World Wide Web Conference*, 1998, http://www7.scu.edu.au/ programme/fullpapers/1834/com1834.htm

11 Gruber, T. Toward principles for the design of ontologies used for knowledge sharing, *International Journal of Human-Computer Studies*, special issue on Formal Ontology in Conceptual Analysis and Knowledge Representation (guest editors: N. Guarino and R. Poli). 1993, http://WWW-KSL-SVC.stanford.edu:5915/doc /network-services.html

12 Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L. and Riedl, J. GROUPLENS: Applying Collaborative Filtering to Usenet News, *Communications of the ACM* 40, 3, 1997, 77-87.

13 Lenz, M. and Glintschert, A. On Texts, Cases, and Concepts, to appear in *Proceedings XPS-99*, Springer Verlag, LNAI.

14 Lowder, J. and Wu, X Wide area selection as a hyperdocument search interface, in *7th International World Wide Web Conference*, 1998, http://mamba.cis.fu-berlin.de:8080/www7/1854/com1854.htm

15 Maeda, H., T. Hirata T. and Nishida, T. COMEMO: Constructing and Sharing Everyday Memory, in *Proceedings of the Ninth International Conference on Tools with Artificial Intelligence* (ICTAI'97), 1997, 23-30.

16 Maher, M.L., Balachandran, M.B. Zhang, D.M. *Case-Based Reasoning in Design*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1995.

17 Nonaka, I., Takeuchi, H., *The Knowledge Creating Company: How Japanese Companies Create the Dynamics of Innovation*, Oxford University Press, New York, 1995.

18 Nakata, K. Voss, A. Juhnke, M. and Kreifelts, T. Collaborative Concept Extraction from Documents, in U. Reimer (ed.), *Proc. Second International Conference on Practical Aspects of Knowledge Management (PAKM 98)* (Basel, Switzerland, Oct 1998).

19 OKAPI Special edition of the *Journal of Documentation* Volume 3, Issue 1, January 1997, http://web.cs.city.ac.uk/research/cisr/okapi/okapi.html

20 Oxman, R. Precedents in design: a computational model for the organization of precedent knowledge, *Design Studies 15*, 2, 1994, 141-157.

21 Prinz, W. NESSIE: An Awareness Environment for Cooperative Settings, to appear *in Proceedings of ECSCW'99* (Copenhagen, Denmark, Sep 1999), http://orgwis.gmd.de/projects/nessie/

22 Resnick, P. and Varian, H. Recommender systems, *Comm. ACM* 40, 3, 1997, 56-58.

23 Röscheisen, M., Winograd, T. and Paepcke, A. Content Ratings and Other Third-Party Value-Added Information Defining an Enabling Platform, *D-Lib Magazine*, August 1995, http://www.cnri.reston.va.us/ home/dlib/august95/stanford/08roscheisen.html

24 Strauss A. and Corbin, J. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*, SAGE Publications, Newbury Park, CA, 1990.

25 Voss, A. and Kreifelts, T. SOAP: Providing people with useful information, in S. C. Hayne, W. Prinz (eds.) *Proc. GROUP'97*, Int. ACM SIGGROUP Conf. on Supporting Group Work - The Integration Challenge (Phoenix AZ, Nov 1997), ACM, New York NY, 291-298.