Collaborative Concept Extraction from Documents

Keiichi Nakata, Angi Voss, Marcus Juhnke and Thomas Kreifelts
German National Research Center for Information Technology
GMD-FIT.CSCW
Schloß Birlinghoven, 53754 Sankt Augustin, Germany
Keiichi.Nakata@gmd.de

Abstract

A group of individuals who share the same interest or a task, would profit from making use of the knowledge possessed by the group. It is then essential that such a body of knowledge, or "community knowledge", be captured in an effective manner. This paper describes the notion of Concept Index, which aims to index important concepts described in a collection of documents belonging to a group, and provide user-friendly cross-references among them to aid concept-oriented document space navigation. Unlike approaches relying primarily on automatic concept extraction tools, the Concept Index relies on users in identifying important concepts by marking keywords and phrases that interest them. Once the concepts are extracted in this manner, they are then enhanced by automated tools and more importantly by users who inspect them. We argue that with an appropriate support for users provided by a system, this interactive process optimises the index generated, and enhances collaboration between the members of the group in managing acquired information, and furthermore, leads to by-products such as a set of community vocabulary that are essential to efficient organisational work.

1 Introduction

It is widely accepted that the management of knowledge held by a group of people sharing a common interest or task, such as a community or an organisation, is an essential aspect of efficient groupwork. Such "community knowledge" is not a simple aggregation of knowledge held by individuals in the group; it also includes knowledge about the existence of knowledge, and competence of the group as a whole. Therefore, it should naturally involve a collaborative process through which the knowledge about knowledge is captured and maintained. This is our primary interest in the context of knowledge management.

The context of this research is the investigation of the concept of "Social Web", an infrastructure that facilitates Internet-based social activities such as collaborative work and forming groups of people with similar interests [GMD98]. The notion of "community knowledge" is one of the central themes in this framework. Similar concepts of network-based social communities have also been put forward in "Sociable Web" by Donath and Robertson [DR94] and in "Knowledgeable Community" by Nishida [Nis95].

In this paper, we describe the notion of a *Concept Index* which intends to capture community knowledge based on collections of documents attributed to that community, and a prototype tool to support its generation and maintenance. In this research paradigm, we base our work on the following assumptions.

- *Documents communicate knowledge.* Documents contain concepts and their relations, which can be seen as the building block of knowledge. Knowledge is coded in documents, and in collections of related documents. Documents can be produced by an individual or collaboratively by a group. Documents are the primary means of asynchronous information exchange, conveying and communicating knowledge. This form of knowledge communication can be supported by making

---

Proc. of the 2nd Int. Conf. on Practical Aspects of Knowledge Management (PAKM98)
Basel, Switzerland, 29-30 Oct. 1998, (U. Reimer ed.)
http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-13/

explicit the concepts and concept relations in documents.

- *Community languages evolve through interaction.* The *vocabulary problem*, the problem associated with people using different terms for the same concept has been addressed in HCI [FLGD87]. A community can be characterised by the language they use. Such a language consists of terms whose meanings are agreed upon by the community, and the concepts to which they refer. These languages may be interactively and collaboratively specified by the members of the community, or may be implicitly defined by their usage. Concepts expressed by such a language are grounded in documents produced or accepted in a community. Furthermore, community languages evolve through acquisition of new information and interactions among the members. Facilitating the inspection of term usage can enhance this process and concepts are grounded in documents.

- *Collections of documents form emergent body of knowledge.* While a collection of documents can be treated as the product of a collaborative effort by a group to document acquired and generated knowledge, analysing the relations between documents provides an implicit, emergent body of knowledge. That is, by making explicit the knowledge contained in documents, group members are made aware of the body of knowledge that exists, together with its relation among each other. In other words, "meta" knowledge is engendered by this process, which can be seen as an emergent property of the collection of documents.

We first describe the Concept Index, claim its benefits, and proceed to discuss the importance of its interactive and collaborative aspect.

2 Concept Index

The motivation for devising a Concept Index is to capture relations between documents as relations between the concepts described, referred to or discussed in these documents. This is intended to capture the knowledge "stored" in these documents, and more importantly, it has the potential to support the emergence of new knowledge by identifying concept relations, making these explicit and enabling users to inspect and edit these concept relations.

A Concept Index is generated in three steps, or levels: *lexical*, *semantic* and *pragmatic* levels. These levels progressively provide richer concept models as well as increased user involvement in the management of the Concept Index. It is important to note that users can use and edit Concept Indexes at any level: indeed the generation of a Concept Index itself is a collaborative task.

2.1 Lexical level: generation of index and cross-references

In this first step, words and phrases that describe the important concepts introduced in the document are identified, extracted, and indexed with cross-references for given collection of documents specified by the user. Since an index is constructed for such a collection of documents, we refer to the documents in the collection as *registered* documents for that index. The author or the reader of a document, by the use of *keyword tags*, specifies occurrences of words in the document that are to be included in the index (vocabulary). This can be performed by highlighting those words in the document, analogous to underlining or using highlighter to mark printed documents. This is an active behaviour on the side of the author, the information provider, and the reader, the information recommender, with the intention of providing a set of concepts which is to be included in his or her contribution to the shared information base. A keyword tag, despite its name, can be assigned to a word, a phrase (a sequence of words), or a URL. Hereafter in this paper the term *keyword* is used with this feature in mind, which means it may consist of any number of words. At this stage, we say that keywords are *exported* from documents to an index, producing a *lexical* index for the registered documents.

The set of index entries is the union of the sets of keywords exported from all the registered documents. The registered documents are cross-referenced and this is reflected as tags in every document. This means that it is possible that the words and phrases that you as a member did not originally mark as keywords are tagged, as a result of cross-referencing based on the lexical index that includes exported keywords from other members of the group for the same document, and those from other registered documents. We refer to the type of keywords that are introduced in this way as of the type *imported.* Therefore, when you read the same document the next time, you will see the keywords you highlighted together with those marked as important by other members of the group. Since these imported keywords may have been introduced from other registered documents via the index, by following these tags, readers can jump from one document to another, or view several documents side by side regarding the parts of those documents that refers to a particular word. This would provide support for reuse of documents, consistent use of terms, identifying relations between documents, and automatic enhancement of a product document.

Although it is encouraged to identify keywords and people would often do so when compelled (for instance, specifying keywords for submission of papers for publication and generating indexes for a document using a word processor), this is not always essential; even when authors or readers do not, for one reason or other, specify keywords themselves, if necessary, we would be able to use term extraction techniques to assist this process [FH96, Gre98]. However, the significance of user-specified keywords is discussed later in the paper. The use of an index is not only useful for users to navigate through the space of registered documents, but also extends to the examination of external documents. Given any document, we would be able to identify and import keywords defined in a selected index that appear in that document. For example, if we come across a new Web page, we might be able to see how important that page is according to the view represented in the chosen index. In this manner, such an index provides a new tool to document exploration.

2.2 Semantic level: enhancement by word relations

Since each keyword would have synonyms, the procedure above can be extended to include synonyms for an indexed keyword. For example, if *doctor* and *GP* both appear in the lexical index, these words should be cross-referenced between documents since they are synonyms. Moreover, if a user introduces another synonym *physician* into the index, a document containing this word can be searched and cross-referenced. The vocabulary is therefore first extended by synonyms of the keywords at this level.

The introduction of synonyms is not only useful to enrich the vocabulary; it can also help distinguish between different meanings of words. For example, for the same word *particle*, there could be two synsets {*particle*, *molecule*} and {*particle*, *function word*}, which represent different concepts.

At the same time, to capture more of the conceptual modelling aspect of a Concept Index, in addition to synonyms, hierarchical relations between concepts, both in terms of inheritance (super/subclasses) and mereological relations (part-of/has-part), can be exploited to provide more enriched relations between documents. Here the vocabulary is extended to include the concepts in hierarchical and compositional relations, and again these are used for cross-referencing.

Information concerning synonyms and word relations can be obtained automatically using a tool such as WordNet [MBF$^+$90], but users should also be able to suggest these relations as well. It is the users' decision to import or not to import the results from these tools.

Through treatment of synonyms and conceptual relations, the index is enhanced onto the semantic level. To reiterate the collaborative aspect of the evolution of Concept Indexes, users are expected to inspect the generated index at any stage. Therefore, automatic disambiguation of polysemy (words that belongs to several concepts) and treatment of synonymy (concepts described by several words) is not considered, since these should be identified interactively by users.

2.3 Pragmatic level: enhancement by concept relations

We can further enhance a Concept Index by identifying related concepts. This can be performed by a user who specifies that two or more concepts are related, or mechanically by identifying co-occurring words in a document, by applying, for instance, text mining techniques [FD95, IBM98]. Text mining is an application of data mining to texts. It is based on statistical methods and is capable of identifying words that co-occur frequently in a given collection of texts. Text mining can identify concepts that are co-related, i.e., those which may not be related in the sense of conceptual relations, but are related in a specific context, thereby generating contextually related concepts. Again, it is the users' decision to use this type of tool.

While it is debatable whether such statistical relations bear significance in terms of conceptual relations in a pure sense, it definitely cannot be dismissed as irrelevant. For example, as the result of text mining on a collection of documents on world trade analysis, a strong co-occurrence between "Korea" and "electronic goods" is found. Korea and electronic goods have no conceptual relation that can be found by WordNet. However, anyone who inspects the result within the context of world trade would immediately relate these; there definitely is a conceptual relation. In fact, this kind of relations is also captured by AI knowledge representation schemes such as semantic nets and conceptual graphs, and have been developed further by recent efforts in building ontologies [Gru93]. From that point of view, inputs from ontology databases such as Ontosaurus [Ont98] can be exploited, but this is beyond the scope of Concept Index at this stage.

What is identified as relations between concepts at this pragmatic level is rather arbitrary. Moreover, these relations, those specified by the user and those identified by text mining alike, are represented as a bag relation "related" in the concept description. However, we believe that such arbitrary relations are often sufficient for identifying related documents, and this increases the expressiveness of Concept Index.

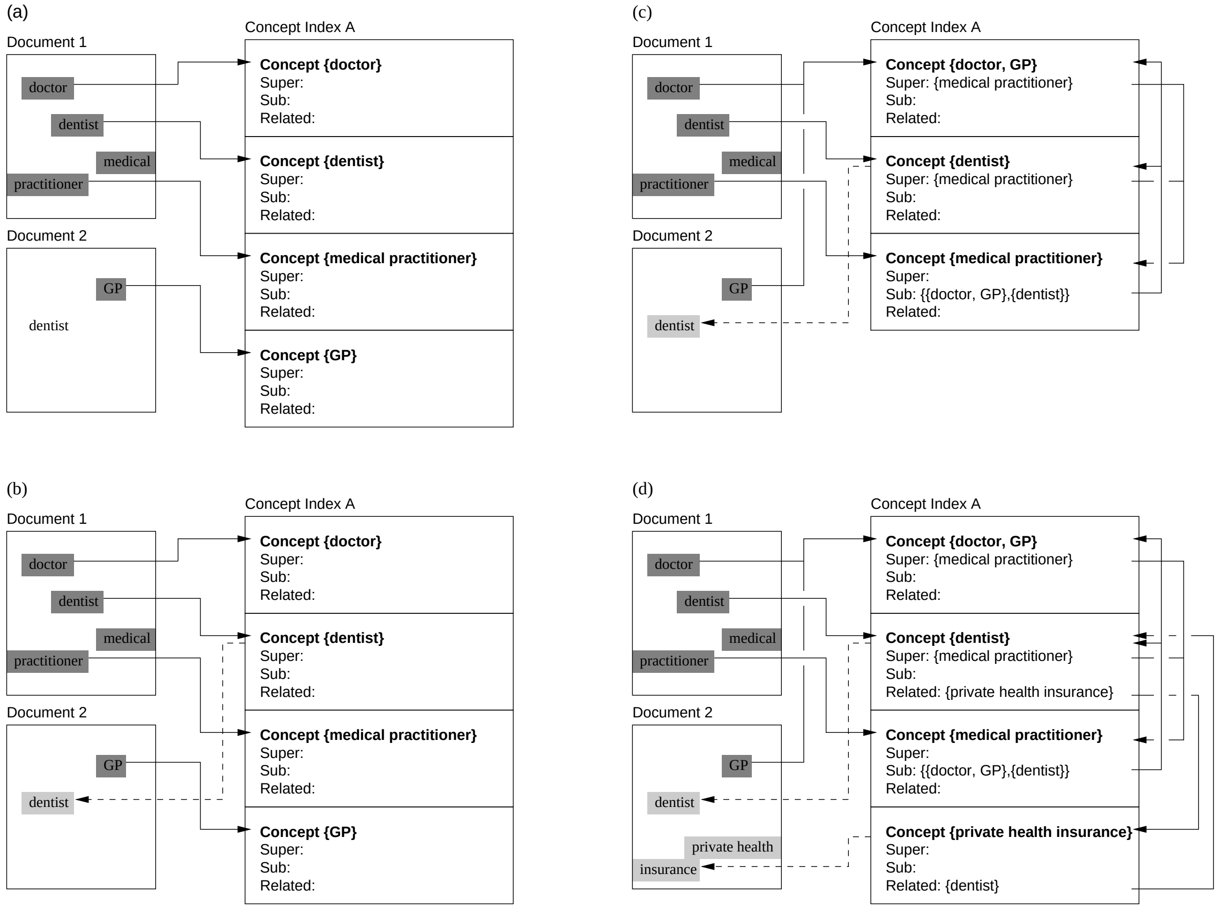


Figure 1: Registered documents and a Concept Index. First, all the marked keywords are inserted into the Concept Index as new concepts ("exported") (a). Then the occurrences of the concepts in other documents are identified ("imported") (b). In (c), synonyms *doctor* and *GP* are merged into one concept, and relations are specified. A relation to a new concept can be inserted, which in turn is imported into documents (d).

2.4 Collaborative editing of the Concept Index

It is an important requirement of a representation of community knowledge that group members would be able to edit the index and remove/add words at each level. To cater for this requirement, one of the external representations of Concept Indexes should be in the form of a document, which is controlled under a collaborative work environment. Since a Concept Index is essentially a semantic net [Qui68], it can be represented in terms of nodes (concepts) and arcs (connections between concepts) as in the example above. However, it can also be seen as a relational database which stores links between concepts and documents. In this view, we can have two ways of interacting with the Concept Index. One is direct interaction, in which users would view and edit the index itself as a document via an editor that acts as a front-end to the index database. The other is by storing and updating documents that contain keywords marked by the user, i.e., users would indirectly interact with the index. Here we first look at the latter form of interaction.

In this form of interaction, Concept Indexes act as a mediator between documents. This means there is no direct, hard-coded reference from a keyword in a document to that in another document, but only through a lookup procedure on a Concept Index, with the exception of explicit hyperlinks from one document to the other. For events concerning storing and viewing of documents the cross-referencing operations are invoked as follows:

1. *Register a document with a Concept Index.* This can be done implicitly by placing the document in a specific folder for which the index is maintained, or explicitly by registering it with a certain index.
2. *Storing a new document.* The first step is the extraction of user-identified keywords. If the keyword already exists in the Concept Index as an

entry of a concept, a link is generated between the occurrence of the keyword in the document and the concept. Otherwise, a new concept is generated and its occurrences in other registered documents are identified. These operations are referred to as "exporting" keywords. Then the document is scanned to find keywords of the concepts that are stored in the Concept Index. If there is any, the occurrence of keyword in the document is added to the concept. This operation is referred to as "importing" keywords.

3. *Viewing a document.* Actual cross-references are inserted to the document to enable document space navigation (see item 5 below). If the document has not been indexed before (e.g., a document before registration, viewing external documents), then keywords need to be imported.
4. *Update/save document.* The reference links between the Concept Index and the document are updated. This may include deletion of a document from the list of registered documents (deregistration, i.e., removal of links to the document).
5. *Document space navigation.* The registered documents can be inspected by traversing links, i.e., following relevant tags via the Concept Index.

In this scenario, the Concept Index serves as the database for links between keyword occurrences in documents.

To illustrate how Concept Indexes are generated, consider the following simple example involving just two documents and a few concepts in conjunction with Figure 1 (part-whole relation entries are omitted for simplicity). Documents 1 and 2 are registered with Concept Index A. Document 1 contains keywords marked by the user, *doctor*, *dentist* and *medical practitioner*, and Document 2 contains *GP*. These are all introduced, or "exported", into Concept Index A as individual concepts, creating concepts {*doctor*}, {*dentist*}, {*medical practitioner*} and {*GP*} (Figure 1(a)). The documents are then scanned for cross-references. The word *dentist* is found in Document 2, so the concept {*dentist*} is linked, or "imported", to this occurrence (Figure 1(b)). These operations are at the lexical level.

In Figure 1(c), the concepts {*doctor*} and {*GP*} are identified as synonyms, by the user or by a thesaurus, and these two concepts are merged into a single concept {*doctor*, *GP*}. In addition, a user or a thesaurus identify {*medical practitioner*} as a super-concept of the concepts {*doctor*, *GP*} and {*dentist*}, and these relations are updated in the index. These are semantic level operations.

At the pragmatic level, as the result of a text mining process, phrases *dentist* and *private health insurance* are identified as having a strong co-occurrence (or a user has claimed that these are strongly related). A new concept {*private health insurance*} is created, which has a "related" link to the concept {*dentist*} (Figure 1(d)).

2.5 Benefits of the Concept Index

The following lists the benefits of the Concept Index primarily in the context of collaborative work.

- *"On-the-fly" concept identification.* The collaborative construction and evolution of a shared Concept Index requires minimal effort. A Concept Index can be developed in a distributive way by both authors and readers tagging important phrases in documents. To ameliorate cold start or support enhancement of one's own vocabulary, it can potentially be initialised or enriched by other sources, such as a thesaurus, shared ontology, classification scheme, and existing Concept Indexes. This procedure, together with text mining tools and concept enhancement features provide a rapid identification of concepts in documents.
- *"Concept spotting".* Since existing concepts are "tagged" in the documents upon browsing and indicated by different colours and their cross-references are generated, readers are provided with visual cues concerning keywords that are likely to reflect relevant issues. Furthermore, in a Concept Index, concepts are described by a set of synonymous phrases, and the existence of any of these phrases are detected and suggested as an occurrence of the concept. Thus, viewing a document with a Concept Index detects occurrences of concepts, rather than phrases.
- *Concept-based infrastructure for document space navigation.* Since cross-references between documents based on related concepts are automatically generated and maintained by the Concept Index via hyperlinks, it will be simple to navigate through the registered documents, and the users are freed from the overhead of manually creating links. Alternatively, users can navigate through the concept structure and explore the document space from concepts.
- *Live cross-reference links.* All references are automatically kept up-to-date, and links are generated upon the inspection of documents. This means that the links are newer than documents themselves.

- *Presentation of documents from different perspectives.* A Concept Index can be viewed to embody a shared language of a community of users, and it can be used to explore the important concepts in the community. Using different Concept Indexes can offer different viewpoints in inspecting the same document.
- *Standardisation of term usage.* Since the maintenance of a group-oriented Concept Index would involve discussions among group members over (dis)agreements concerning the usage of a term for a concept, it would contribute to standardising the term usage within a group. A similar effect could be obtained by navigating the document using concepts, since users would be made aware what terms are commonly used by other documents to describe the same concept. This is in line with the approach taken by WebOnto [Dom98] in building ontologies.

We have described the potential benefits of the Concept Index. One of the main features of the Concept Index is that most of the processes in its generation and maintenance involves user interaction. In the next section we discuss the significance of this user involvement.

3 User-oriented concept extraction

There have been serious efforts in the area of natural language processing to automatically extract concepts in documents by means of linguistic methods based on language parsing. However, despite these efforts this approach still lack robustness and more importantly has serious performance problem. Commercial systems such as Agentware [Aut98] have hence opted for statistical approaches. Automatic concept extraction, if successful, provides an easy way of generating indexes.

The overhead of concept extraction by the user, who essentially identifies keywords themselves, is perhaps a laborious (to the user) aspect of the Concept Index approach. At first glance, this seems to be a case of abandoning automation due to the complexity of the task and limitations in automated tools, and simply reverting back to mundane human labour. However, we argue that there are benefits, rather than drawbacks, to this approach.

3.1 Specification of central concepts from author and readers' points of view

First of all, the specification of keywords by the author and the readers constitutes an intentional act of communicating to other potential readers what they find as important in the document. This is akin to an author specifying keywords that suitably describe the concepts central to the document, or using emphatic fonts such as italic or bold face for significant words and phrases to provide visual cues to the reader. Such an overhead is not exceeding for an author who has a vested interest in making his or her composition more readily accessible to readers with specific interest in the topic, as seen in conference and journal paper submissions and in the insertion of meta-information into HTML headers. At the same time, the highlighting of text corresponds to the action of a reader who would use highlight markers to mark words that he or she regards as important or interesting to be noted for future reference (see Section 3.3) or to draw attention of others who would read it later.

Automatic concept extraction is useful when there is no intention on the author and reader's side to communicate information to a specific audience or a targeted group of people. It would use statistical means to process the documents and extract concepts mainly based on noun phrase extraction, and find correlations among them. However, we argue that in a collaborative situation, such a passive participation in collaboration merely hampers efficient information exchange.

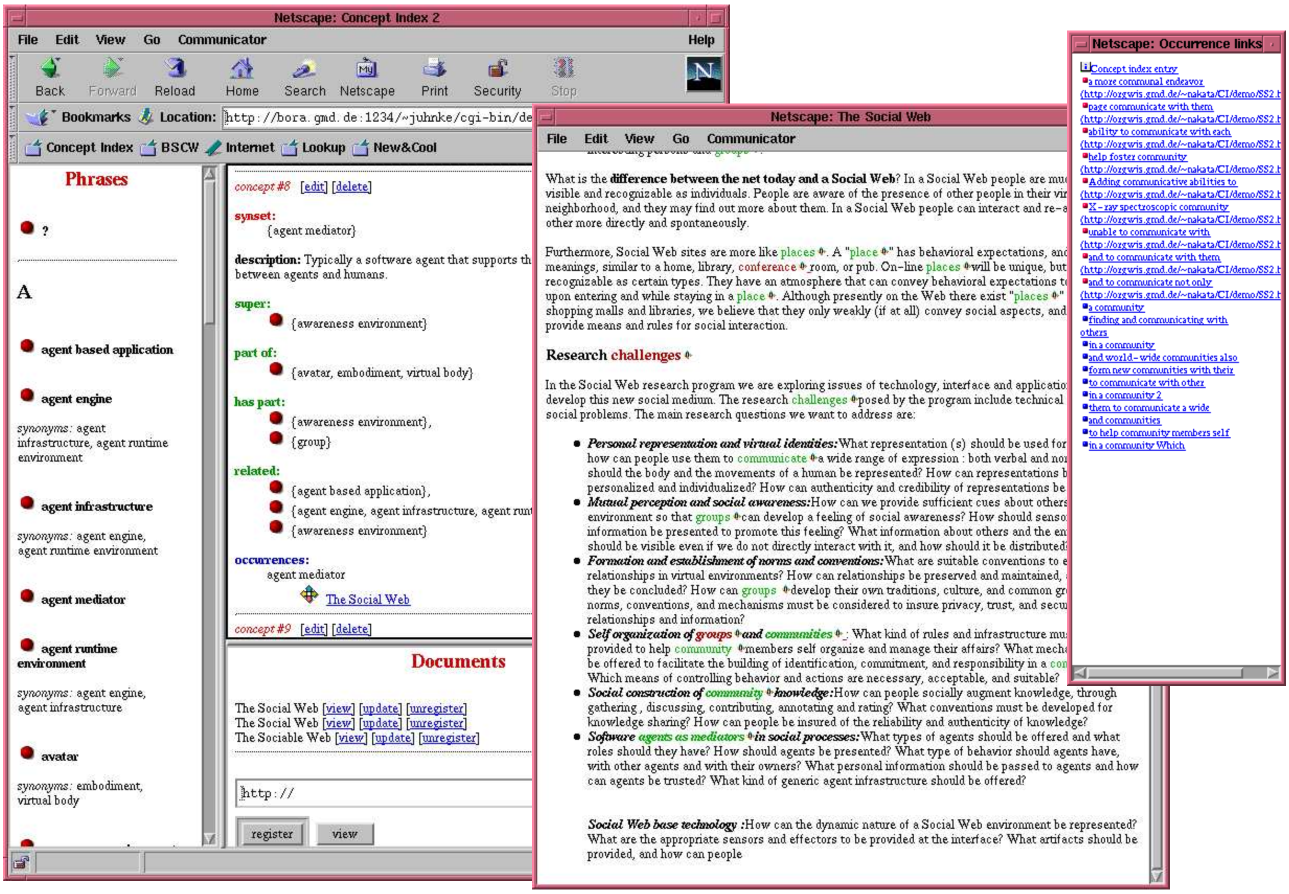
3.2 Index size optimisation

Another aspect of user specification of concepts is that identified concepts would reflect the interests of the group to which the user belongs. The active role the users play in identifying important concepts provide exactly this feature. At the same time, this optimises the size of the index to the necessary minimum, and avoids it being cluttered with irrelevant concepts.

Clearly, automatic concept extraction is effective and the Concept Index can be enhanced by the results it can produce. Furthermore, even without user intervention, automatic concept extraction can also filter out concepts that are not relevant to the user group [CMNS96] However, the way to do so is to use a profile such as a group thesaurus or a term list to remove concepts that do not appear in them. This has the same result as when the Concept Index is used, but there is an additional overhead of constructing user or group profiles. The approach taken in the Concept Index effectively combines these two processes into one, by reversing the process, i.e., instead of filtering the machine generated results, it enhances the user specification of concepts by an automatic tool.

3.3 Collective memory

The highlighting of parts of text constitutes another conscious action by the user. People often highlight or underline words so that they can easily spot the



http://bora.gmd.de:1234/~juhnke/cgi-bin/de

Phrases

● ?

A

● **agent based application**

● **agent engine**

synonyms: agent infrastructure, agent runtime environment

● **agent infrastructure**

synonyms: agent engine, agent runtime environment

● **agent mediator**

● **agent runtime environment**

synonyms: agent engine, agent infrastructure

● **avatar**

synonyms: embodiment, virtual body

concept #8 [edit] [delete]

synset:
{agent mediator}

description: Typically a software agent that supports th
between agents and humans.

super:
● {awareness environment}

part of:
● {avatar, embodiment, virtual body}

has part:
● {awareness environment},
● {group}

related:
● {agent based application},
● {agent engine, agent infrastructure, agent runt
● {awareness environment}

occurences:
agent mediator
The Social Web

concept #9 [edit] [delete]

Documents

The Social Web [view] [update] [unregister]
The Social Web [view] [update] [unregister]
The Sociable Web [view] [update] [unregister]

http://

What is the **difference between the net today and a Social Web**? In a Social Web people are muc
visible and recognizable as individuals. People are aware of the presence of other people in their vir
neighborhood, and they may find out more about them. In a Social Web people can interact and re-a
other more directly and spontaneously.

Furthermore, Social Web sites are more like places. A "place" has behavioral expectations, and
meanings, similar to a home, library, conference room, or pub. On-line places will be unique, but
recognizable as certain types. They have an atmosphere that can convey behavioral expectations to
upon entering and while staying in a place. Although presently on the Web there exist "places"
shopping malls and libraries, we believe that they only weakly (if at all) convey social aspects, and
provide means and rules for social interaction.

Research challenges

In the Social Web research program we are exploring issues of technology, interface and applicatio
develop this new social medium. The research challenges posed by the program include technical
social problems. The main research questions we want to address are:

- ***Personal representation and virtual identities:*** What representation (s) should be used for
how can people use them to communicate a wide range of expression : both verbal and non
should the body and the movements of a human be represented? How can representations b
personalized and individualized? How can authenticity and credibility of representations be
- ***Mutual perception and social awareness:*** How can we provide sufficient cues about others
environment so that groups can develop a feeling of social awareness? How should senso
information be presented to promote this feeling? What information about others and the en
should be visible even if we do not directly interact with it, and how should it be distributed
- ***Formation and establishment of norms and conventions:*** What are suitable conventions to e
relationships in virtual environments? How can relationships be preserved and maintained,
they be concluded? How can groups develop their own traditions, culture, and common gr
norms, conventions, and mechanisms must be considered to insure privacy, trust, and secu
relationships and information?
- ***Self organization of groups and communities*** _: What kind of rules and infrastructure mu
provided to help community members self organize and manage their affairs? What mech
be offered to facilitate the building of identification, commitment, and responsibility in a co
Which means of controlling behavior and actions are necessary, acceptable, and suitable?
- ***Social construction of community knowledge:*** How can people socially augment knowledge, through gathering , discussing, contributing, annotating and rating? What conventions must be developed for knowledge sharing? How can people be insured of the reliability and authenticity of knowledge?
- ***Software agents as mediators in social processes:*** What types of agents should be offered and what roles should they have? How should agents be presented? What type of behavior should agents have, with other agents and with their owners? What personal information should be passed to agents and how can agents be trusted? What kind of generic agent infrastructure should be offered?

Social Web base technology : How can the dynamic nature of a Social Web environment be represented? What are the appropriate sensors and effectors to be provided at the interface? What artifacts should be provided, and how can people

Concept index entry
a more communal endeavor
(http://orgwis.gmd.de/~nakata/CI/demo/SS2.h
page communicate with them
(http://orgwis.gmd.de/~nakata/CI/demo/SS2.h
ability to communicate with each
(http://orgwis.gmd.de/~nakata/CI/demo/SS2.h
help foster community
(http://orgwis.gmd.de/~nakata/CI/demo/SS2.h
Adding communicative abilities to
(http://orgwis.gmd.de/~nakata/CI/demo/SS2.h
X-ray spectroscopic community
(http://orgwis.gmd.de/~nakata/CI/demo/SS2.h
unable to communicate with
(http://orgwis.gmd.de/~nakata/CI/demo/SS2.h
and to communicate with them
(http://orgwis.gmd.de/~nakata/CI/demo/SS2.h
and to communicate not only
(http://orgwis.gmd.de/~nakata/CI/demo/SS2.h
a community
finding and communicating with others
in a community
and world-wide communities also
form new communities with their
to communicate with other
in a community 2
them to communicate a wide
and communities
to help community members self
in a community Which

Figure 2: The prototype Concept Index interface. Each concept entry can be viewed and edited through the index browser interface (*left*). A document viewed using the index (*centre*) displays exported concepts (red in colour display) and imported concepts (green), each followed by an icon indicating that the phrase appears in the index. When the icon is clicked, links to the concept entry and the occurrences of the concept in other places among registered documents appears in the link window (*right*).

components which they found interesting at a later point in time. Since the Concept Index retains all the words and phrases that are marked by keyword tags, this information is never lost until someone makes a conscious decision to delete it. Furthermore, a user can annotate the entry in the Concept Index.

This kind of feature already exists in commercial word processors and in Web page annotation [DB98, Cha98]. Annotation of a portion of a text can be seen as an effective means of communication, by pinpointing to the location of discussion in the text. It can also be, more interestingly, a rather casual act by the reader to note what is so interesting about it, so that he or she can come back to it later. This action, repeated by other members in the same group, generate what can be called as "collective memory". That is, a piece of memory gathered and combined by a group of people. To fulfil its role as memory, users should be able to inspect it, use it to find out whether it is worth remembering it, etc., making it more "personal" to the group.

Obviously, automatic concept extraction is not intended for this kind of task, and it is not constructive to overstate this feature. However, the point here is that in the Concept Index, there is no distinction between concept extraction and annotation. These are treated the same which makes it more dynamic and specific to the group. The expressiveness of loosely structured relations of information items can be seen in work by [MKA+97, NHM97].

4 Prototype

Currently, we are developing a prototype system that constructs and maintains Concept Indexes. In this implementation, Concept Indexes are implemented as relational database, which stores concept entries with synsets, links to super/sub concepts, links for part-whole relations and other relations, and links to the keywords in the documents. All the documents are treated as HTML documents and keywords can be tagged using a specific colour or introducing a new tag. When keywords are exported, they are first "normalised", i.e., converted into singular nouns and stripped of any verb inflections (for every word if the keyword consists of several words; stopwords are also removed) and stored into the database. If the keyword to be exported already exists in the index, only the reference to the concept in the index is created; otherwise, a new concept is created together with the reference. When documents are scanned to identify keywords that should be imported, each word is stemmed to improve matching with entries in the database (again, stopwords are ignored). If the document contains keywords stored in the index, then tags are inserted and the reference to the concept is created. Figure 2 shows the screendumps from the prototype system interface.

Note that the process above is purely mechanical. This means, for instance, when an exported keyword is assigned to more than one concept due to polysemy, all possible references are created which might contain inappropriate references. A similar process occurs in importing keywords. For this reason, the index itself needs to be inspected by the user for maintenance.

5 Issues and future prospects

5.1 Issues

From the experience of using the prototype system, we address three issues which need to be resolved in the further development of the system. These are the problem with the performance of the system, necessity of further linguistic processing, and the way the indexes are managed.

- *Performance.* The operations involved in constructing and maintaining the Concept Index is inherently complex, since the process of importing (i.e., identifying concepts in the Concept Index appearing the text) requires thorough text processing and database lookup. Therefore, the complexity is increased by the number of documents in the collection multiplied by that of concepts in the index. Furthermore, this operation is repeated every time the Concept Index is updated. However, since the performance is paramount to this type of system, this is a serious limitation that would hamper any attempt to scale up, and efforts are being made to improve the performance. One non-algorithmic solution is to use agents as background processes to constantly monitor the changes in the Concept Index [VGH$^+$98].
- *Further linguistic processing.* In the current implementation, keywords (including phrases) are spotted in the text by matching phrases in the synset. For example, if *doctor* and *GP* are treated as describing the same concept forming a synset, then every occurrence of *doctor* and *GP* is cross-referenced. Since the stemmed form of the word is used for matching, *modification* would match *modifier.* Also, when matching phrases, they are matched within a reasonable proximity, so the phrase in the Concept Index *open architecture* would match with *architecture with robustness and openness* in the text. However, when the phrases become more complex, there could be a need for further linguistic processing and concept extraction. For example, the phrase *efficient implementation* should perhaps match with *fast program*, since *fast* and *efficient* can be seen as synonyms and so are *implementation* and *program.* However, this will involve breaking down every phrase into single-word concepts and the higher possibility of matching both increases the complexity of the Concept Index and the computation involved. In addition, there is no guarantee that such an operation would result in reasonable matches and there is a danger of proliferation of unsuitable matches between phrases.
- *Distributed versus central index management.* While the assumption is that a group would maintain its own Concept Index, there are two ways of storing it. One is to have a central index, which has the advantage in the ease of maintaining consistency of data but has the disadvantage of being possibly too large, both for inspection and performance. The other is to have small, even personal, indexes which can be merged into one large index whenever necessary. This would have the pros and cons opposite to the first approach. This issue also requires experimentation and is not easily resolved. However, it is clear that we should be able to merge indexes and inspect or use indexes through a filter such as contributors and subsets of registered documents.
- *Towards ontologies?* One of the motivations for the current design of Concept Index was to reduce the overhead of generating a prescribed knowledge base. Unless there is an overwhelming benefit that justifies the effort of constructing an ontology, it is unlikely that any common user would contribute to that activity. While efforts are made to alleviate the effort of describing an ontology, such as in Ontobroker [FER98] and WebOnto [Dom98], it is a task that requires some determination. Concept Indexes are not intended to build ontologies, although we would benefit from importing results from existing ones. Concept Indexes

aim to support specifications of less rigorous relations, or associations, which might be more intuitive to contributors and lead to more interesting links via associations (such an observation is made in [NHM97]). However, once constructed, ontologies are powerful sources of knowledge. It is worth considering whether the Concept Index approach can contribute to the process of ontology construction.

The issues described above cannot be immediately resolved and require further experimentation and feedback from the user community.

5.2 Future prospects

We plan to increase the functionality of the system to enhance its capability, and the following are two of such extensions that are being considered.

- *Marking and matching a larger portion of text.* So far we have been dealing with words and phrases, but in theory there is no limit to the number of words in the phrase. This means in some situation, users might wish to mark sentences or even paragraphs. In fact, we have seen cases where readers highlight the complete paragraph, annotating it with comments. Clearly there is a functional difference between a phrase and a sentence or paragraph, and while a phrase is likely to describe a concept, a sentence or paragraph would describe more than one concept and possibly even their relations. One solution is to instruct the user to limit the marking to a sequence of words describing only one concept. However, this increases the cognitive load on the user, and two users may not agree on what constitutes a concept. To further complicate the issue, if we were to offer the functionality such as importing concepts for a sentence or paragraph, then this should be matched with a corresponding portion of text in other documents.
- *Indexing images.* In the area of image processing, attempts have been made in the indexing of images (for example [SC96]). However, as in the case of automatic concept extraction, they are mostly research systems and lack generality and robustness [Eak96]. However, if images can be tagged by words that describe them, then it can also be indexed by the Concept Index. This offers extension from text only documents to those incorporating images as their essential component. A promising approach is to take the same stance as in user-oriented concept extraction and expect users to “mark” the important images and label them with words. Such an approach is taken by [CIDT97].

It is worth reiterating that tools of this kind are only meaningful if they are extensively used by a community of users. One of the top items in our agenda is to perform experiments in real-life situations to test the feasibility of this approach and identify users’ needs. Any extension of functionality naturally takes this requirement into consideration.

6 Conclusion

The Concept Index is intended to support the process of collaborative concept extraction and management. Concepts are extracted by means of highlighting words and phrases in a document as keywords that represents important concepts described in the document. This conscious act of concept identification is performed by any member of the group that shares the same interests and it can be seen as an implicit mode of communication since the extracted concepts are indexed, cross-referenced in related documents and used to navigate the document space. Furthermore, a Concept Index can be edited by the members of the group, and this leads to collaborative development of “community language” which is essential in efficient groupwork.

We have argued that this collaborative process of both concept extraction and management has a number of advantages over fully automatic concept extraction. The main argument is that the conscious effort of identifying keywords enhances collaborative work and optimises such an index to a necessary minimal. We have developed a prototype system, and we intend to carry out further experiments to test the feasibility of this approach and refine user requirements.

References

[Aut98] Autonomy, 1998. http://www.agentware.com.

[Cha98] B.-W. Chang. In-place editing of web pages: sparrow community shared documents. *Computer Networks and ISDN Systems*, 30:489–498, 1998.

[CIDT97] K. Chandrinos, J. Immerkær, M. Dörr, and P. Trahanias. A visual tagging technique for annotating large-volume multimedia databases. In *Fifth DELOS Workshop Filtering and Collaborative Filtering*, pages 125–129, Budapest, 1997.

[CMNS96] H. Chen, J. Martinez, T. D. Ng, and B. R. Schatz. A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment on the worm community system. *Journal* *of the American Society for Information Science*, 47(8), 1996.

[IBM98] IBM Corporation. IBM intelligent miner for text, 1998. http://www.software.ibm.com/data/iminer/foretext/tatools.html.

[DB98] D. DaLiberte and A. Braverman. A protocol for scalable group and public annotations, 1998. http://www.hypernews.org/l̃iberte/www/scalable-annotations.html.

[Dom98] J. Domingue. Tadzebao and WebOnto: Discussing, Browsing, and Editing Ontologies on the Web. In *11th Knowledge Acquisition for Knowledge-Based Systems Workshop*, Banff, Canada, April 1998.

[DR94] J. Donath and N. Robertson. The sociable web. In *Proceedings of the Second International World-Wide-Web Conference*, 1994.

[Eak96] J. P. Eakins. Automatic image content retrieval: are we getting anywhere? In *Proceedings of the third International Conference on electronic Library and Visual Information Research*, pages 123–135, De Montfort University, Milton Keynes, 1996.

[FD95] R. Feldman, I. Dagan. Knowledge Discovery in Textual Databases (KDT). In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, pages 112–117, Menlo Park, CA, 1995. AAAI Press.

[FH96] R. Feldman, H. Hirsh. Mining Associations in Text in the Presence of Background Knowledge. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 343–346, Menlo Park, CA, 1996. AAAI Press.

[FER98] D. Fensel, M. Erdmann, and R. Studer. Ontobroker: The Very High Idea. In *Proceedings of the 11th International FLAIRS Conference (FLAIRS-98)*, Sanibal Island, Florida, May 1998.

[FLGD87] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

[GMD98] GMD-FIT.CSCW. The social web. new forms of interaction in virtual environments: A research framework, 1998. http://orgwis.gmd.de/projects/SocialWeb.

[Gre98] S. J. Green. Automated link generation: Can we do better than term repetition? *Computer Networks and ISDN Systems*, 30:75–84, 1998.

[Gru93] T. Gruber. Toward principles for the design of ontologies used for knowledge sharing. Technical Report KSL 93-04, Stanford University Knowledge Systems Laboratory, 1993.

[MBF^{+}90] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.

[MKA^{+}97] H. Maeda, M. Kajihara, H. Adachi, A. Sawada, H. Takeda, and T. Nishida. Weak information structures for community information sharing. *International Journal of Knowledge-Based Intelligent Engineering Systems*, 1(4):225–234, 1997.

[Nis95] T. Nishida. The knowledgeable community: Towards knowledge level communication. In *The Seventh International Forum on the Frontier of Telecommunications Technology*, Tokyo, Japan, November 1995.

[NHM97] T. Nishida, T. Hirata, and H. Maeda. CoMeMo-Community: a system for supporting community knowledge evolution. In *Proceedings of the First Kyoto Meeting on Social Interaction and Communityware*, Kyoto, Japan, June 1998.

[Ont98] Loom Ontosaurus. Loom web browser, 1998. University of Southern California, Information Sciences Institute, Intelligent Systems Division. http://www.isi.edu/isd/ontosaurus.html.

[Qui68] M. R. Quillilan. Semantic memory. In M. Minsky, editor, *Semantic Information Processing*. MIT Press, 1968.

[SC96] J. R. Smith and S.-F. Chang. Tools and techniques for color image retrieval. In *IS&T/SPIE Symposium on Electronic Imaging: Science and Technology–Storage & Retrieval for Image and Video Databases IV*, 1996.

[VGH+98] A. Voss, H. Guo, H.-L. Hausen, M. Juhnke, T. Kreifelts, K. Nakata, and V. Paulsen. Agents for collaborative information exploration. In *Proceedings of the Third International Workshop on CSCW in Design (CSCWID'98)*, Tokyo, Japan, July 1998.