Title: APPARATUS, METHOD AND SOFTWARE FOR GENERATING A KNOWLEDGE PROFILE AND THE SEARCH FOR CORRESPONDING KNOWLEDGE PROFILES

Abstract: The invention regards a method, an apparatus and software for generating a knowledge profile from textual information, in which textual information is compared to at least one structured datafile, per word in the textual information all concepts from the structured datafile related to it are coupled, subsequently concepts are clustered into a list of concepts or covering concepts by means of clustering, after which the list of concepts is interactively presented to a user. The invention also relates to a method for building up and maintaining a knowledge and/or interest network.
Published: — without international search report and to be republished upon receipt of that report

For two-letter codes and other abbreviations, refer to the “Guidance Notes on Codes and Abbreviations” appearing at the beginning of each regular issue of the PCT Gazette.
Apparatus, method and software for generating a knowledge profile and the search for corresponding knowledge profiles.

The invention relates to a method for generating a knowledge profile from textual information, as well as an apparatus and software for it. The invention additionally relates to a method, an apparatus and software for searching for corresponding knowledge profiles.

Such methods, apparatus and software are generally known. It is for instance known to analyse, catalog or summarize textual information by means of so-called "natural language processing". The correlation between words is analyzed on the basis of current grammar rules, that depend on the language, and in that way relations between words that together form concepts are recognized and linked. Such an analysis however takes a lot of computing time, is complex and depends on the language.

Additionally methods, apparatus and software are known in which all words in a text, except fillers, are indexed one by one. However, a lot of information is lost in the process, especially with regard to concepts that are compiled of several words. Additionally the knowledge profiles obtained become large. In such systems generally only searches in a boolean manner can take place.

Other known methods, such as among others described in US-5.931.907, US-5.754.938, EP-A-860.785 or WO-A-0/17781, use keywords. A drawback then is that using the wrong keywords when searching leads to missing out on information. Additionally a keyword can be used in for
instance a document that has nothing to do with documents looked for. For instance the use of the keyword "xenotransplant" during a search, may lead to missing out on references in which the word "xenographic procedure" is used. Additionally truncation may lead to the search term "xeno" and thus to too many irrelevant hits.

In yet other known methods, such as among others described in WO-A-98/38560, use automatically generated word clusters and terms. In such methods correlation and relations between words are recognized by processing very many texts. When certain words occur together often, these words may be recognized as belonging to one concept.

As a result, the known methods are either too slow, or too inaccurate, to for instance, be used interactively and maybe by an inexperienced user.

It is an object of the invention to provide a method of the kind mentioned in the preamble which is among others suitable for interactive use, for use by an inexperienced user, or for use in a distributed environment, such as for instance the internet or an intranet.

These objects are achieved, and other advantages are gained, with the method for generating a knowledge profile from textual information, in which:
- textual information is entered into a computer;
- the words of the textual information are looked up in at least one structured datafile, which structured datafile comprises words and for each of those words a reference to at least one validated concept belonging to that word;
- per word found in the structured datafile a list of words is generated containing all validated concepts related to it from the structured datafile;
- subsequently by means of clustering the validated concepts from the
various lists of words are clustered into the knowledge profile
comprising a list of concepts or covering concepts.

The invention further relates to an apparatus for generating a knowledge
profile from textual information, comprising a computer provided with
connection means for realising data transfer with other computers, in
which the computer comprises:
- first memory means for storing a structured datafile provided with a
  list of words and references to validated concepts;
- second memory means for storing the textual information;
- a data processing unit and third memory means with software
  provided with search routines for looking up words from the textual
  information in the list of words of the structured datafile, extraction
  routines for extracting from said list of words representations for
  validated concepts in lists of words per word and with cluster
  routines for reducing the lists of words to a clustered list.

Additionally the invention relates to software for generating knowledge
profiles in which the software is provided with search routines for looking
up the words from the textual information in the list of words of the
structured datafile, extraction routines for extracting from said list of words
representations for validated concepts in lists of words per word, with
cluster routines for reducing the lists of words to a clustered list.

Because the method, the apparatus and the software according to the
invention in the generating of knowledge profiles uses specific structured
datafiles, and considers all words in a text to be processed, it is possible to
process textual information in a quick, efficient manner which may even be
suitable for interactive use by an inexperienced user, for instance via a
distributed environment such as the internet or an intranet.

In this text, cataloguing among others means defining the knowledge con-
tents of a piece of textual information by means of a knowledge profile, and storing knowledge profiles in a catalog, preferably called a collection. The textual information is provided here with a range of validated concepts, that are a part of the knowledge profile. This indexing, which means attributing a list of validated concepts, of textual information renders the information easily and accurately to be found. The validated concepts that are linked to the textual information preferably consist of abstract representations, such as numbers. One representation therefore stands for a validated concept. The validated concept may be a description of one or several words.

A knowledge profile may in addition be based on one single piece of textual information. Additionally the knowledge profile may also be based on several pieces of textual information. This in fact makes a knowledge profile of one single piece of textual information a concept profile, and a knowledge profile, for instance based on all publications of a research institution, a knowledge profile. However, in this text the concept knowledge profile is used for both.

In this text the term computer does not only relate to a single, so-called "stand-alone" computer, but also to an arrangement of computers, that are interconnected with means to exchange information. The computers may each separately carry out the method according to the invention or be provided with software according to the invention. It is also possible that each computer carries out a part of the method or is provided with a part of the software. The computers may be organised in a fixed network, but also be connected to for instance the internet or an intranet. The computers may connected to each other through a physical line, but wireless as well.

The computers may also be organised like a client-server system, or in a peer-to-peer organisation.
An example of an abstract representation of a concept among others is the International Patent Classification system (IPC), in which concepts, often consisting of several words, are indicated by means of a code. The awarding of the IPC code as usual however is entirely different from the way in which according the invention the representations are awarded. In the classic use of the IPC code a text is first interpreted, preferably by a human reader. When the meaning and the contents are known, the suitable IPC codes are found by means of the tree structure of the IPC code. In contrast to the invention, the IPC for instance additionally uses carefully chosen keywords. The invention specifically uses indices or representations representing validated concepts. A fingerprint is thus made of a part of textual information. That means that with the help of the words in the text and a structured datafile, a list of validated concepts representing the meaning of the text is searched for.

Preferably the representations are numbers. As a result the computer can quickly compare lists of representations to each other and cluster them.

By a validated lists of concepts is meant a list of concepts preferably compiled by people, preferably experts. A compilation of validated concepts thus is an ontology representing a field of knowledge or a piece of knowledge. In for instance medical science a disease may have many different names. By selecting a representation for a specific disease, and all different words for that disease, but also letting all words that together describe that disease, refer to that representation, it is prevented that by selecting one name the document cannot be found back when using another name, or that other articles using another name are not found. The invention is among others based on the insight that a multitude of in themselves ambivalent words, due to the fact that they occur together in a piece of text, particularly when they occur in each other’s proximity, may represent a very clearly defined concept.
In this text the term "word" is used as the smallest unit in a text of roman characters that could represent a concept, like for instance "chest". In other languages, such as Japanese or Chinese, a word may correspond to one character.

An additional advantage of selecting a representation for a concept, is that the possibility is created to work independent of language. For instance a knowledge profile can be generated from an English text, in a list in a second language, for instance French, the corresponding concepts can be searched for and represented, and in a file of knowledge profiles and references to knowledge sources in a second or third language the corresponding knowledge sources in that language can be searched for by means of the representations, and subsequently be presented to a user. The knowledge profiles of said found knowledge sources can be presented in the second language, so that the user gets an impression of the contents of the knowledge sources without consulting the knowledge sources in their original language.

Preferably the structured datafile comprises a structured list of words such as a thesaurus or meta-thesaurus. When the word thesaurus is used in this text, meta-thesaurus is also implicitly meant. In thesauri concepts are classified according to a hierarchic system of covering or generic concepts with below them each time more specific concepts. This results in a kind of tree-structure of higher, covering concepts, each time branching to more specific concepts. In the method according to the invention, if so desired, use can be made of thesauri of various knowledge fields and said thesauri can be combined into one large thesaurus.

Preferably clustering takes place based on relationships within sentences. In this way a very fast method is possible, whereas it appeared from tests that still a very good accuracy was achieved. In the clustering common concepts are searched for and possible corresponding covering concepts to
which words in a sentence refer back. This is repeated again and again until no more common concepts are found.

Clustering can be carried out by running a window that is two words wide over the text and searching for commonly validated concepts in the two lists, subsequently running a window that is three words wide over the text and comparing the three lists until a maximum width of the window.

In order to be able to determine which concepts are going to be shown to a user in the list of validated concepts, all validated concepts in the list of validated concepts are provided with weights that indicate the importance with regard to each other. Preferably the weights comprise quantities regarding the frequency with which the concepts occur in the textual information, the specificity of the concepts and a measure for the certainty of the concept occurring (i.e. the sensitivity) in the textual information. The weights indicate statistic characteristics of each concept. This among others includes the specificity, the sensitivity, the number of alternatives occurring in the textual information and the textual similarity. On the basis of the weights it can for instance be determined which concepts from the list of concepts are shown to the user. When for instance many words in a text all refer to a long list of concepts, but all these lists contain among others the same concept, then it will be likely that that concept is meant in the text. When in addition it appears that these words, which among others refer to the same concept, are in close proximity of each other in the text, the likelihood that that concept is described in the text increases even further. An example of this is a text in which the (English) term "black water fever" occurs. "Black" may for instance refer to the concepts "colour", "race", but also to "malaria". When in the lists of references of both "water" and "fever" the concept "malaria" also occurs, it is very likely that the disease "malaria" is meant.

To increase the accuracy even further it is preferred when the user, after
presentation of the generated list of concepts with accompanying weights, is able to interactively adjust the weights, specifically the relative weights of the concepts shown to him with respect to each other. To render the textual information easy to find back, information with regard to the location of the textual information is included in the list of concepts. As a result the list of concepts with the location can be included in a datafile, and the textual information can easily be found back. By including references the complete textual information need moreover not be included. As a result it is easy to build up a structured datafile from unstructured information, or a distributed datafile is build up in a simple way. Preferably the information relating to the location is a hyperlink or URL or another reference to the information in computers that are connected or can be connected by means of a network or in another way. This results in a distributed datafile, in which the textual information can even be distributed over very many different computers. An example of this are computers that are connected or can be connected through the internet.

Another way to increase the accuracy is by adjusting the weights of each validated concept of an input knowledge profile for which a knowledge source in a catalog of knowledge profiles and knowledge sources coupled to it have to be searched, to the degree to which the validated concept in the input knowledge profile is specific with respect to the knowledge profile in the catalog. When for instance a search takes place in a catalog containing only knowledge sources relating to for instance malaria, a concept such as "malaria" will not be specific to that catalog, and the weight of that concept in the input knowledge profile will be decreased.

The invention further relates to a method for cataloguing textual information, in which the textual information is compared with at least one structured datafile, per word in the textual information all concepts from the structured datafile that are related to it are coupled, subsequently the concepts are clustered into a list of concepts or covering concepts by
means of clustering, after which a list of concepts is interactively presented to a user.

The invention further relates to a method for building up and maintaining knowledge and/or interest networks, in which textual information is catalogued according to the method for cataloguing textual information as described above, and in which the list of concepts are coupled to information for identifying the user, preferably a hyperlink or E-mail address.

Additionally the invention relates to a method for searching in textual datafiles, in which textual input is catalogued according to the method for cataloguing textual information as described in the above, after which a knowledge profile is searched for in the textual datafile showing the largest similarity to the list of concepts to the list of concepts containing the formalised query.

The method according to the invention appears to be very suitable for use in interactive environments and interactive uses, and in particular for interactive internet or intranet uses. For interactive uses, and more specifically for, for instance, internet and intranet uses, the speed and the quantity of data that several computers, for instance a server and the personal computer or terminal of a user, have to exchange with each other to come to a wanted result, are of importance.

It is possible for a user to process a second piece of textual information after analysis of a first part of textual information. The two lists resulting from it are subsequently combined into one list by combining the concepts in the list on the basis of the weights.

A possible use for which the method according to the invention is particularly suitable, is the interactive building up and maintenance of knowledge and/or interest networks ("communities"), particularly through
the intranet or internet. Software for carrying out the method according to the invention is present on a server. A user can approach the software on the server. After entering personal data, the user is enabled to transfer textual files selected by him to the server, for instance by means of "drag-and-drop" or via a "file transfer". This may be files written by him, such as for instance a curriculum vitae, but better yet a longer text, such as reports, an extended essay, a dissertation, articles or the like. In that case a knowledge profile is created. Textual files may also be articles that represent the field of interest of the user. In that case it regards an interest profile. By storing the interest profiles or knowledge profiles of very many persons and making them searchable, for instance by storage in a datafile, a knowledge or interest network is created. It is also possible for a person to create knowledge profiles from various texts and to combine those knowledge profiles.

The concepts may also be addresses of knowledge sources such as knowledge institutes, universities, companies, government institutions or experts in a field of study. When addresses are scattered in a piece of textual information, the method according to the invention will refer each part of the address to the same knowledge source, as a result of which a complete address can be generated, possibly in a structured format.

In a computer the software (that means the software searches for representations for the concepts) indexes the textual information in accordance with the method according to the invention and presents a list of concepts to the user. The user is subsequently enabled to adjust the list, for instance by changing the weights awarded per concept. This adjustment can take place in various interactive ways. For instance use can be made of spider's web diagrams. In these diagrams various concepts are radially ranged around a common centre point. By dragging a concept along a radial axis, for instance by means of input means such as a mouse, keys, a trackball, a touch pad or the like, the relative weight of a selected
concept can be changed. Another possibility is to plot the concepts on a bar chart, and by means of the input means already mentioned enabling the user to set the length of the various bars. After that the user can store the list of concepts and connections with its textual information locally in his own computer or server. The list can, if so desired, be fed and added to a larger datafile of data of other users.

With the help of the list of concepts and their weights, a user can also search in a datafile. This searching can take place interactively. The user then for instance sees the number of hits. By interactively changing the weight of the various concepts, for instance in a manner described above, the user immediately sees the number of hits change.

In the method according to the invention, the words in the structured datafile preferably refer to representations for validated concepts, and the list of words and the clustered list of representations comprise the validated concepts. This renders it possible to work independent of language. Additionally very high processing speeds can be achieved on a computer, and a large data compression can take place, as a result of which the method can be carried out by means of for instance a light server.

It is preferred in a method according to the invention that the knowledge profile comprises a list of representations of validated concepts.

Preferably the knowledge profile is interactively presented to a user.

Preferably the structured datafile comprises a thesaurus or meta-thesaurus, in which each word in the thesaurus or meta-thesaurus is provided with a reference to a validated concept. As a result existing systems can be connected to without much effort.
Preferably clustering takes place based on relationships within sentences.

More preferably or instead clustering takes place by means of comparing lists of words of representations of words in close proximity of each other, based on words within sentences, or based on words within a pre-defined distance from each other.

Preferably clustering comprises the following steps:

- comparing lists of words of validated concepts of words standing next to each other in the textual information;
- selecting concepts these words have in common;
- subsequently comparing lists of words of validated concepts of words which are further apart from each other in the textual information and again selecting concepts that such words have in common;
- compiling the list of validated concepts of the knowledge profile from the selected concepts;
- awarding a weight to each of the validated concepts in the knowledge profile, in which each validated concept is awarded a higher value as the distance between the words is smaller.

In the method according to the invention the concepts in the list of concepts of a knowledge profile preferably are provided with weights that indicate the importance of the various concepts with regard to each other.

In this way a so-called fingerprint of textual information is created. Said fingerprint appears to be very unique for the text. To such an extent even that from tests it appeared that when said fingerprints match for over 80%, it regards one and the same text, or plagiarism, this despite the fact that in plagiarism the order of sequence often is different and synonyms of words in the original are used.

Preferably the weights are related to the frequency of words in the textual
information referring to the concept. Additionally or instead the weights may be related to the specificity of the concepts. Additionally or instead the weights may provide a measure for the likelihood of the concept occurring in the textual information. Preferably the user can adjust the weights, preferably interactively.

In the knowledge profile preferably information is included relating to the location of the textual information, preferably in the form of a hyperlink. In this way the possibility is offered to find the original information back quickly, without storing entire documents in a computer. A distributed system can be build up in that way.

To make quick processing possible, the words in the structured datafile preferably are normalised words and the textual information preferably is converted into a list of normalised words, after which the normalised words are looked up in the structured datafile.

Additionally the invention relates to a method for building up and maintaining knowledge and/or interest networks, in which a user enters textual information into a computer, after which a knowledge profile is generated from the textual information in accordance with the method according to the invention, and in which the knowledge profile either information or a reference is included for identifying the user, preferably a hyperlink or E-mail address. In this way the knowledge becomes controllable, manageable and searchable, even at light, remote systems.

Preferably the generated knowledge profile is presented to the user by means of display means connected to the computer, after which the user can remove concepts from the knowledge profile and add concepts to the knowledge using input means, connected to the computer. Preferably the user can adjust the weights in the knowledge profile.
The invention additionally relates to a method for searching for knowledge sources, in which textual information is entered into a computer, a first knowledge profile is generated from the textual information in accordance with the method according to the invention, the computer is provided with a datafile provided with knowledge profiles and references belonging to each knowledge profile, which references refer to the knowledge source from which the knowledge profile has been compiled, in which the computer is provided with software provided with search routines to search for at least one knowledge source having a knowledge profile showing similarities according to pre-entered threshold values with the generated first knowledge profile.

In this way quick and accurate searching is possible.

Preferably the software is provided with selection routines for selecting those knowledge profiles which within limits statistically show the largest similarity to the generated first knowledge profile.

The invention further relates to an apparatus for generating a knowledge profile from textual information, comprising a computer comprising input means, output means, memory means, data processing means and connection means for connection to other computers, in which the computer is adapted for carrying out the method according to the invention.

Additionally the invention relates to a method for generating the knowledge profile of a text or text excerpt, in which a user enters at least one text or text excerpt into the computer, in which the computer is provided with software with which consecutively
- a list of words is being generated from the text or text excerpt;
- these words are being looked up in at least one structured list of words, present in the computer memory, in which the structured list of words comprises words and for each word in the structured list of
words references to representations for validated concepts with which that word may be connected;
- in the computer memory lists of representations are being generated from the structured list of words;
- the lists of representations are being clustered into one clustered list in the computer memory, the clustered list comprising representations representing the knowledge profile of the text or text excerpt.

Preferably the software awards a weight to each representation in the clustered list.

The invention additionally relates to a method for searching for knowledge sources, in which a user enters at least one text or text excerpt into a computer, in which the computer is provided with software with which consecutively
- a list of words is being generated from the text or the text excerpt;
- said words are being looked up in at least one structured list of words, in which the structured list of words comprises words and for each word in the structured list of words references to representations for validated concepts with which that word may be connected;
- lists of representations are being generated from the structured list of words;
- the lists of representations are being clustered into one list of representations representing the knowledge profile of the text or the text excerpt;
- in a datafile provided with references to knowledge sources and concepts connected to said knowledge sources, those knowledge sources are being searched for, of which the concepts coupled to it show maximum similarity to the list of concepts in the clustered list;
- a list of references to the knowledge sources, of which the coupled
The invention additionally relates to software for generating a knowledge profile from a text, in which the software is provided with routines for carrying out the above-mentioned method.

The invention additionally relates to software for searching for knowledge sources, in which the software is provided with routines for carrying out the above-mentioned method.

A specific embodiment of the invention will be elucidated on the basis of the figures. The figures are illustrations of one or more embodiments of the invention, and should not be seen as a limitation of them or to that end.

Figure 1 shows a schematic view of a specific embodiment of the method according to the invention.

Figure 2 shows the relation between identification data of users.

Figure 3 shows a possible way for building up a knowledge and/or interest network.

Figure 4 shows a possible construction of a structured datafile according to the invention.

Figures 5 and 6 show a view of an embodiment of the method for generating a knowledge profile according to the invention.

Figure 7 shows a user interface in which the software according to the invention is used.
Figure 8 shows a possible presentation of a knowledge profile.

Figure 1 shows an example of an implementation of the method according to the invention. A text excerpt, or an entire article, or a series of articles are entered (1). This entering can for instance take place by the user marking text parts or selecting text files in the user’s computer by means of input means such as the mouse, and dragging the selected text files or marked text parts to an input screen that is shown on the user’s computer screen by the software on the server.

A text excerpt or text file, the textual information, is subsequently first normalised (2). The fillers are then removed (3), and words are reduced to their stem ("stemming"). This results in a list (4) of normalised words. Said list is then compared (5) to concepts in a thesaurus or a meta-thesaurus (6). By using an alphabetical list of words from the thesaurus or meta-thesaurus derived from the thesaurus or meta-thesaurus that have been reduced to their stem, list of stem words, it appeared to be possible to let comparison take place very quickly. Preferably the stem words have been ordered in an n-ary tree to make it possible to quickly look up the stem words in the list of stem words with each word in the list of normalised words. All possible thesaurus concepts are searched for for each word in the list of normalised words in the textual information. In this way a list (7) is produced with all possible thesaurus concepts for each word of the textual information. Words that do not occur in the thesaurus are left out.

Subsequently an analysis of the results (8) takes place. For each word per sentence, it is established whether there is a relation with another word in the sentence, i.e. whether two or more words together are a part of a concept that occurs in the thesaurus or meta-thesaurus. The top-layer thesaurus concepts are also searched for and, when several bottom-layer concepts refer to a same top-layer concept, are replaced by the top-layer concept. This process is called clustering. Clustering is the search for common concepts that are referred back to in the text by adjacent words.
The clustering is again applied on clusters found, until no common concepts are found. If so desired clustering can take place within sentences first. That means that it is established whether common concepts occur. This can be repeated until no change occurs any more. After that clustering on the basis of adjacent sentences may take place. After that it can be analysed whether there are concepts that are a specimen of the same top-layer concept, i.e. have a common genus. In that case the genus can be inserted. Possibly that can also only be done for the concepts that are presented to the user.

Each concept found is provided with a weight. This weight is among others compiled from a value indicating where a concept is located on the scale of specific to general. These values have been given beforehand to the concepts in the thesaurus. Additionally the weight is compiled from the frequency with which the concept occurs in the textual information. The weight is further compiled from a likelihood number that indicates how certain the software is that the concept corresponds to the words in the textual information. On the basis of the weights it is determined which concepts in the list of concepts are presented to the user. Its selection criterion can be set.

Subsequently a list of suggested concepts (9) is presented to the user. The user is then enabled to interactively adjust the list (10). The user can then return the adjusted list (11). In the adjusted list a connection (12) is included to the original textual information, for instance in the form of a hyperlink to the text, possibly in another computer, an address or E-mail address of the author or another user, or in another way. Preferably however, a connection to the textual information in another computer. As a result not many data need to be stored on the server. Because of that a relatively light server will suffice, as a result of which the computer system according to the invention can be designed light.
The method according to the invention can very advantageously be deployed in the development, maintenance and build-up of knowledge and interest networks of persons within organisations, of organisation-to-organisation and/or person-to-person. Figures 2 and 3 relate to that. To build up and maintain such a network, knowledge and interest profiles of persons and organisations have to be generated and connected to each other. The method according to the invention can support and implement this.

In figure 2 it is schematically shown which information parts have to be entered to that end, and what the interrelations between the various information parts are. It is for instance possible the include data regarding persons (20) such as the name, with which organisation they are employed, an E-mail address and other data. Additionally, data regarding the organisation (21) can be included, such as contact data, but also an interest or knowledge profile (22). This profile can be generated by means of the method according to the invention. Additionally or instead a knowledge or interest profile of the person (23) can be included, with connections to textual information (24). This interest or knowledge profile can be generated by means of the method according to the invention.

Figure 3 shows a possibility on which a method for building up and maintaining a knowledge network according to the invention can be implemented. A user (31) first of all enters textual information (sources) which according to the user relate to his expertise, such as articles and reports written by him, or his interests, such as articles directly relating to the field of interest. Said textual information is processed in accordance with the method according to the invention, for instance in accordance with the chart of figure 1. From this processing follows a knowledge or interest profile, coupled to the user data and to the sources. The user adjusts the profile interactively (32). Subsequently the profile is queued (33). A authorization unit (34), either being an automated system or a
person, checks the data and the profile for completeness and carries out a validation, before entering the data and the profile in a datafile, i.e. a database (35). When entering in the database the user receives a confirmation message (36). The database can be consulted by users.

In figure 4 the generating of a structured datafile according to the invention from an index, such as UMLS is shown. In UMLS the concept "Schilder's disease" has "C0014044" as representation. Various terms belong to it, such as among others "diffuse sclerosis". In the structured datafile according to the invention, each separate word belonging to the representation "C0014044", so for instance "disease" and "sclerosis", is normalised and put in a separate list of words, with a link to the representation "C14044".

In figure 5 a example is shown of the method according to the invention, in this case with an English text. Fillers are removed from the sentence, and the remaining words are normalised. The representations of the concepts to which this word refers are subsequently looked up for each word in the structured datafile. In this example 527 validated concepts were found with "assessment". The representations stand on the left-hand side of the list, behind that, for the sake of clarity, the textual meaning of the validated concept. Representation 11741 here stands for the validated concept "determination of health care quality".

In figure 6 an example of clustering is shown. A window, in this case two words wide, is run over a sentence and search takes place in the lists of (representations of) validated concepts whether words in a window have corresponding validated concepts. In this case, here in this figure on the basis of English words, the word "assessment" leads to the concept indicated with the representation "C220825", and which has the English description "evaluation". The words "efficacy" and "drug" lead to 47 concepts.
In figure 7 a user interface is shown, in this case of an internet or intranet application. In the field "content text" the text can be entered from which a knowledge profile has to be assembled, to which a search may have to be made. This entering in a control system like Windows can for instance take place via delecting in a word processor and transfer by means of "drag-and-drop". The piece of text has been processed in this example, the validated concepts found are in the field "IKA Terms", as well as their respective weights, indicated here also by means of the length of the bars next to the validated concepts.

By means of an input means, in this case a mouse, a user is enabled to adjust the length of the bars, resulting in a change of the relative weight of a validated concept. During search this adjustment of the weights remains interactively possible, and because the method allows a very quick implementation in the software, a user almost instantaneously sees the number of knowledge sources found change. Such an application shown in the figure, can for instance be deployed to search within a yellow pages.

Figure 8 shows another possible presentation of a knowledge profile. In this case the validated concepts are indicated like dots of a spider’s web, in which the distance to the centre point corresponds to the weight.

The distance to said centre point can be interactively adjusted here.

Additionally the invention relates to what is known as "knowledge mining". With the method according to the invention it namely is also possible to very quickly carry out knowledge mining on a textual file or files, by making a knowledge profile or knowledge profiles and combining them.

The invention additionally regards a method in which validated concepts in a list of concepts are connected by also validated semantic relations from the structured datafile on the basis of the textual information, as a result of
which semantic networks on the level of concepts are created.

In this case semantic relations between concepts are stored in the structured datafile as separate entities and representations.
Claims

1. Method for generating a knowledge profile from textual information, in which:
   - textual information is entered into a computer;
   - the words of the textual information are looked up in at least one structured datafile, which structured datafile comprises words and for each of those words a reference to at least one validated concept belonging to that word;
   - per word found in the structured datafile a list of words is generated containing all validated concepts related to it from the structured datafile;
   - subsequently by means of clustering the validated concepts from the various lists of words are clustered into the knowledge profile comprising a list of concepts or covering concepts.

2. Method according to claim 1, in which the words in the structured datafile refer to representations for validated concepts, and the lists of words and the clustered list comprise representations of the validated concepts.

3. Method according to claim 1 or 2, in which the knowledge profile comprises a list of representations of validated concepts.

4. Method according to one or more of the preceding claims, in which the knowledge profile is interactively presented to a user.

5. Method according to one or more of the preceding claims, in which the structured datafile comprises a thesaurus or meta-thesaurus, in which each
word in the thesaurus or meta-thesaurus is provided with a reference to a validated concept.

6. Method according to one or more of the preceding claims, in which clustering takes place based on relationships within sentences.

7. Method according to one or more of the preceding claims, in which clustering takes place by means of comparing lists of words of representations of words in close proximity of each other, based on words within sentences, or based on words within a pre-defined distance from each other.

8. Method according to one or more of the preceding claims, in which clustering comprises the following steps:

- comparing lists of words of validated concepts of words standing next to each other in the textual information;

- selecting concepts these words have in common;

- subsequently comparing lists of words of validated concepts of words which are further apart from each other in the textual information and again selecting concepts that such words have in common;

- compiling the list of validated concepts of the knowledge profile from the selected concepts;

- awarding a weight to each of the validated concepts in the knowledge profile, in which each validated concept is awarded a higher value as the distance between the words is smaller.

9. Method according to one or more of the preceding claims, in which the concepts in the list of concepts of a knowledge profile are provided with weights that indicate the importance of the various concepts with regard to each other.
10. Method according to claim 9, in which the weights are related to the frequency of words in the textual information referring to the concept.

11. Method according to claim 9 or 10, in which the weights are related to the specificity of the concepts.

12. Method according to claim 9, 10 or 11, in which the weights provide a measure for the likelihood of the concept occurring in the textual information.

13. Method according any one of the claims 9-12, in which the user can adjust the weights, preferably interactively.

14. Method according to one or more of the preceding claims, in which in the knowledge profile information is included relating to the location of the textual information, preferably in the form of a hyperlink.

15. Method according to one or more of the preceding claims, in which the words in the structured datafile are normalised words and the textual information is converted into a list of normalised words, after which the normalised words are looked up in the structured datafile.

16. Method for building and maintaining knowledge and/or interest networks, in which a user enters textual information into a computer, after which a knowledge profile is generated from the textual information in accordance with the method according to one or more of the preceding claims, and in which in the knowledge profile either information or a reference is included for identifying the user, preferably a hyperlink or E-mail address.

17. Method according to claim 16, in which the generated knowledge profile is presented to the user by means of display means connected to
the computer, after which the user can remove concepts from the knowledge profile and add concepts to the knowledge using input means, connected to the computer.

18. Method according to claim 17, in which the user can adjust the weights in the knowledge profile.

19. Method for searching for knowledge sources, in which textual information is entered into a computer, a first knowledge profile is generated from the textual information in accordance with the method according to one or more of the preceding claims, the computer is provided with a datafile provided with knowledge profiles and references belonging to each knowledge profile, which references refer to the knowledge source from which the knowledge profile has been compiled, in which the computer is provided with software provided with search routines to search for at least one knowledge source having a knowledge profile showing similarities according to pre-entered threshold values with the generated first knowledge profile.

20. Method according to claim 19, in which the software is provided with selection routines for selecting those knowledge profiles which within limits statistically show the largest similarity to the generated first knowledge profile.

21. Device for generating a knowledge profile from textual information, comprising a computer comprising input means, output means, memory means, data processing means and connection means for connection to other computers, in which the computer is adapted for carrying out the method according to one or more of the preceding claims.

22. Device for generating a knowledge profile from textual information, comprising a computer provided with connection means for realising data
transfer with other computers, in which the computer comprises:
- first memory means for storing a structured datafile provided with a
  list of words and references to validated concepts;
- second memory means for storing the textual information;
- a data processing unit and third memory means with software
  provided with search routines for looking up words from the textual
  information in the list of words of the structured datafile, extraction
  routines for extracting from said list of words representations for
  validated concepts in lists of words per word and with cluster
  routines for reducing the lists of words to a clustered list.

23. Software for generating knowledge profiles in which the software is
provided with search routines for looking up the words from the textual
information in the list of words of the structured datafile, extraction
routines for extracting from said list of words representations for validated
concepts in lists of words per word, with cluster routines for reducing the
lists of words to a clustered list.

24. Method for generating the knowledge profile of a text or text excerpt,
in which a user enters at least one text or text excerpt into the computer,
in which the computer is provided with software with which consecutively
- a list of words is being generated from the text or text excerpt;
- these words are being looked up in at least one structured list of
  words, present in the computer memory, in which the structured list
  of words comprises words and for each word in the structured list of
  words references to the representations for validated concepts with
  which that word may be connected;
- in the computer memory lists of representations are being generated
  from the structured list of words;
- the lists of representations are being clustered into one clustered list
  in the computer memory, the clustered list comprising representa-
  tions representing the knowledge profile of the text or text ex-
25. Method according to claim 24, in which the software awards a weight to each representation in the clustered list.

26. Method for searching for knowledge sources, in which a user enters at least one text or text excerpt into a computer, in which the computer is provided with software with which consecutively
- a list of words is being generated from the text or the text excerpt;
- said words are being looked up in at least one structured list of words, in which the structured list of words comprises words and for each word in the structured list of words references to representations for validated concepts with which that word may be connected;
- lists of representations are being generated from the structured list of words;
- the lists of representations are being clustered into one list of representations representing the knowledge profile of the text or the text excerpt;
- in a datafile provided with references to knowledge sources and concepts connected to said knowledge sources, those knowledge sources are being searched for, of which the concepts coupled to it show maximum similarity to the list of concepts in the clustered list;
- a list of references to the knowledge sources, of which the coupled list of concepts show maximum similarity to the list of concepts in the clustered list, is being presented to a user.

27. Software for generating a knowledge profile from a text, in which the software is provided with routines for carrying out the method of claims 24 or 25.

28. Software for searching for knowledge sources, in which the software
29. Carrier provided with software according to any one of the preceding claims 27 or 28.

30. Method for generating a structured textual datafile, in which a knowledge profile is generated from textual information in accordance with the method according to one of the preceding claims, in which a list is generated with representations of concepts, in which the concepts comprise structured textual information portions, after which the structured textual information portions belonging to the representations are being retrieved and are saved in a structured in a file.

31. Device comprising one or more of the characterizing measures described in the description and/or shown in the drawings.

32. Method comprising one or more of the characterizing measures described in the description and/or shown in the drawings.

33. Carrier provided with software for carrying out the method according to one or more of the preceding claims.
The likelihood of vaccination against malaria is not very high unless we can find a live-altersted vaccine that can be spread by mosquitoes because that would take away the necessity to build up capacity to administer the vaccine. The latter is a real problem in areas with weak health service.
FIG. 2
Fig. 4

Reference Knowledge Source Construction

UMLS '99

IKA KS

diffuse sderosis (crisp)
sderosis, sudanophilic cerebral
(ckpian)
Schilder's Disease (read)
Encephalitis periventricular diffusa (mesh)
leukoencephalitis, subacute (ckplian)

C0014044
(ckpian's disease)

Diffuse
C14044...
Sderosis
C14044...
Sudanophilic
C14044...
Cerebral
C14044...
Schild
C14044...
Disease
C14044...
Encephalitis
C14044...
Periventricular...
Fig. 5
Statistical Pattern Matching

remove stopwords

normalization

concept lookup

...assessment of the efficacy of drug treatment using molecular meth

...assessment efficacy drug treatment using molecular meth

...assessment (527) efficacy (9) drug (3086)
**Fig. 6**

**Statistical Pattern Matching**

...assessment of the efficacy of drug treatment using molecular methods.

<table>
<thead>
<tr>
<th>Clustering</th>
<th>Assessment</th>
<th>Efficacy</th>
<th>Drug</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>210415</td>
<td>[Evaluating]</td>
<td>210433</td>
<td>210433</td>
<td>210433</td>
</tr>
</tbody>
</table>

- 210415: Measure reaction to administration, hematopoiesis and etc.
- 210433: Advice about drug treatment
- 210433: Advice about efficacy of drug treatment
- 210433: Advice about side effects of drug treatment
- 210433: Advice to continue with drug treatment
- 210433: Advice to discontinue drug treatment
- 210433: Advice to skip drug treatment
- 210433: Antifungal drug - chronic treatment
- 210433: AID treatment control
- 210433: AID treatment method
- 210433: AID treatment unit
- 210433: AID treatment variable
- 210433: AID treatment confidentiality
- 210433: AID treatment method
- 210433: Brief Office Visit (for the sole purpose of mountain) (I)
- 210433: Cerebral or lateral ventricular (CP-II) peaces, with or
- 210433: Diabetic neuropathy treatment [no drugs here]
- 210433: Determine patient's pain, extent and frequency of drug &
- 210433: Drug abuse treatment
- 210433: Drug for the treatment of giup allergy reaction
- 210433: Drug for the treatment of giup allergy
- 210433: Drug Timestep and Response to Therapy
- 210433: Drug treatment not indicated
- 210433: Drug treatment stopped - medical advice ICD
- 210433: Drug treatment stopped - patient ran out of tablets
- 210433: Drug treatment unit
- 210433: Drug treatment variables
- 210433: Drugs used for the treatment of acute pain
- 210433: Drugs used for the treatment of drug allergy disease
- 210433: Drugs used in the treatment of gastrointestinal pain
- 210433: Hemorrhage - tongue (black)
Fig. 7

Objectives:
- To determine the genomic organization of var genes in P. falciparum.
- To examine the regulation of var gene expression in P. falciparum.
- To examine the diversity of var proteins.
- To clone var gene homologues from P. vivax.
- To examine the genomic organization of var gene homologues in P. vivax.

Approach:
- Generation of a P. vivax YAC library and its annotation into contigs. A representative YAC library for P. vivax will be constructed. The DNA material will be obtained from patients and from primates infected with an isolate of P. vivax.
- Annotation of P. vivax YAC clones into chromosomal contig maps. Chromosomal YAC contig maps will be generated using a 'top down' approach based on markers for different chromosome regions. YAC clones will be screened with oligonucleotide probes from P. vivax.

Checked concepts are used to describe the projects:

- BGA Technology
- DNA
- YAC
- Genes
- Cloning
- Sequencing
- Hybridization
- Peptides
- BLAST
- BLASTX
- BLASTN

Source document

[Image of a web page in an Internet Explorer window]