PERIODICA POLYTECHNICA SER. EL. ENG. VOL. 45, NO. 2, PP. 91–108 (2001)

DOCUMENT RETRIEVAL THROUGH CONCEPT HIERARCHY FORMULATION

Péter SCHÖNHOFEN and Hassan CHARAF

Department of Automation and Applied Informatics
Budapest University of Technology and Economics
H–1521 Budapest, Hungary
Fax: (+36) 14632871, Tel.: (+36) 14632870
e-mail: {schonhofen, hassan}@avalon.aut.bme.hu

Received: August 22, 2001

Abstract

The enormous growth of the Internet and the widespread use of computer systems in general created very large collections of electronic documents, and methods existing so far have proved unable to handle the massive amount of unstructured documents. In this article we discuss a variant of document retrieval, where traditional indexing is augmented by concept hierarchy (composed by observing concept roles in each member of the document collection) in order to improve accuracy. In addition, the document model is introduced as a way to recognize the limits inherent in document retrieval where no complete text understanding is feasible.

Keywords: information retrieval, document summarization, document retrieval, concept hierarchy, document model.

1. Introduction

The enormous growth of the Internet and the widespread use of computer systems in general created very large collections of electronic documents. Methods existing so far have proved unable to handle the massive amount of unstructured documents which, in addition, may cover a fairly wide range of topics. Natural language understanding and knowledge representation systems are not sufficiently effective and accurate to be of practical use and, on the other hand, common keyword based retrieval algorithms are unable to discover even slightly complex semantics buried under the surface of words following each other.

In the present article the focus will be on document retrieval, that is, selecting documents from a document collection which, considering a user formulated topic, are deemed as relevant. Our aim is to find the middle course between complete text understanding and mere collection of keywords in order to determine the topic of each document – we achieve this by regarding documents not as individual entities, but as members of a larger document collection. First, the limits of our capacity to uncover and describe document content should be clearly understood, and in the light of this knowledge, a feasible goal established. Second, the actual method to process documents and whole document collections should be constructed tak-

ing into account the practical time and storage space restrictions. Third, the entire retrieval mechanism and the way the user interacts with the system should be specified.

2. Task

There are several ways one can look for a given piece of information in a knowledge repository. You may either pose questions ('Where is the nearest post office?') or specify the topic you want to know more of ('Tropical plants in South Africa'). The knowledge repository itself may be an unlimited set of World Wide Web pages, images with captions, given number of plain text documents and so on. Finally, the result can be displayed in multiple formats: list of document references, document extracts grouped by location or topic, browsable category hierarchies etc.

Now we make the following assumptions:

- The document collection is very large (possibly containing several million documents); however, the number of members is known and each member document is accessible at any time.
- Documents are written in the same language for which a syntax parser exists; moreover, documents may include simple formatting instructions (paragraph separator, title, emphasizing and so on).
- Topics are sufficiently continuous and representative; that is, for each document there are documents covering similar topics and for each topic there are general as well as specific documents covering it.

The consequence of the first point is that we cannot employ sophisticated natural language understanding methods, since we do not abound either in time or storage space – we must content ourselves with performing only shallow syntax parsing (like in BEARL et al. (1997) and in EVANS and ZHAI (1996)). Likewise, it is obviously out of the question to examine all documents whenever a user submits a query – instead, a representative should be produced for each document, which is a reliable substitute in the retrieval process. Representatives in turn, although efficient with topic based document retrieval, do not facilitate question answering. As to displaying results, because in most cases users refine or modify their queries according to the document list returned by the retrieval process, the result set should constitute an integral part of the interaction.

Having outlined the desired retrieval system, let us examine what are the deficiencies of existing keyword based search engines (see also FALOUTSOS and OARD (1995)):

- Keyword (or term) extraction methods cannot recognize more complex concepts which are described by more than one word, except when the word construct occurs always in the same form.

- They cannot detect the context of a keyword, which then leads to many unrelated documents included in the result set (for example 'boot' can be both a footwear and a computer initialization procedure).
- Documents are processed and evaluated as separate entities instead of taking into consideration their environment, namely the whole document collection – thus unnecessarily losing valuable information (COLE and EKLUND (1996) and COHEN and SINGER (1996) are two attempts to remedy this).

An ideal document representative may be constructed only when knowing the document, the query submitted by the user and the other documents among which the current document has to be evaluated. Since representatives are built only once, queries remain unknown; however, member documents hold a so far unexploited potential: hence the importance of the last point.

3. Idea

The basic idea for the retrieval system proposed here is as follows. Since users want to retrieve documents both related and relevant to a certain topic or concept, documents have to be broken down to concepts as well – thus representatives should be concept lists, only slightly more complicated than keyword lists are in traditional methods. Owing to its central role, as much information has to be gathered about each concept as possible, namely from two sources: one of them is 'document-relative' and the other is 'document-absolute'. The relative knowledge describes what is the role of the concept in the document (document model), while the absolute knowledge defines how the concept relates to other concepts encountered in the whole document collection (concept structure model). Before the introduction of these two models in the next section, the meaning and characteristics of concepts have to be expounded.

Concepts are everything a user may refer to in a query between possible Boolean operators. (For the sake of convenience and simplicity, from now on we assume that queries comprise a sole topic.) For example 'museum', 'conical shape', 'slowly rotating shaft' and 'parts assembly procedure without employing electrical measurements' are all valid concepts. Even 'the first poem Poe wrote after the death of his wife which was published in a major newspaper or magazine' would, at least theoretically, qualify as a concept. There are two factors restricting concept complexity – ability of the syntax parser and lack of reasoning with appropriate knowledge representation. This limitation arises from the fact that the same concept the user has formulated can appear in many different forms in actual documents, and is not always contained in a single sentence. Besides, the same concept may occur multiple times in a document, but again in various ways: abbreviated, as a pronoun, in an altered grammatical structure and so on. As we specify increasingly complex concepts, we are able to identify fewer and fewer occurrences of it, gradually losing its context.

Therefore simplifications should be made at several points while relying on two principles: first, the complexity we retain must be easily and accurately recognizable by the syntax parser, and second, the complexity we relinquish must not impair our ability to distinguish documents.

- Since technical documents, which characterize the majority of document collections, usually describe static knowledge, the temporal dimension may be omitted with impunity, meaning that verbs be stripped from all tense and modal also information. However, temporal relations can have great importance, as topics like 'modify settings after installation' and 'modify settings before installation' may differ considerably.
- Verbs and participles should be treated as equal, because both serve the purpose of refining the meaning of a noun or pronoun construct – and being consistent means that the same is also true for objects. In short, the noun or pronoun kernel can be extended by other nouns (frequent in scientific terminology, such as 'voltage threshold'), adjectives, adverbs, verbs and objects.
- Words qualifying verbs (verb prefixes), words expressing relations between verbs and objects (adverbs) or between dependent clauses (conjunctions) should be either ignored or incorporated in the verb; otherwise the added complexity in concept representation overwhelms the entire document retrieval process. Besides, grammatical constructs are generally highly redundant.

Of course, the concept formation procedures were only roughly outlined above and their actual implementation significantly varies with different languages having different grammatical structures – due to difficulties in parsing, additional simplifications may be needed. In the following, we treat concepts at a much higher level and will differentiate only three classes of them: primary, auxiliary and composite. Primary concepts have meaning in their own, such as 'plate' or 'fast revolution' but auxiliary concepts do not, for example 'cautiously' or 'yellow'. Consequently, composite concepts are made of one or more primary concepts and an arbitrary number of auxiliary concepts. Though at first glance it would seem that the introduction of this classification is unnecessary, because of the presence of possibly imperfect syntax parsing and certain post-processing methods, it will prove useful.

4. Document Model

The document model describes how we can characterize a document using information obtained by shallow syntax parsing and recognition of certain formatting statements embedded in the text. In brief, the document model defines what knowledge we have about a document during retrieval and this limits both the attainable accuracy and the kind of interactive methods we provide to the user. Of course, the document model is organized around concepts: it is a set of concepts having properties and relating to each other.

Let us now examine what information can we extract from a document (after appropriate pre-processing steps, such as stemming, synonym replacement, omission of non-relevant words and so on):

- Document zones. Formatting instructions might outline paragraphs, sections and chapters in the document, which can significantly improve the correct discovery of concept contexts (see LALMAS and RUTHVEN (1997)). However, creating multiple-level structures is not recommended, due to the possible inaccuracy and small zone sizes – single-level zoning seems sufficient.
- Concepts occurring in the text. Initially only primary and auxiliary concepts are recognized, then grammatically connected concepts are merged into composite concepts while retaining the original concepts. This way one concept may form part of another.
- Frequency of concepts. Unfortunately, frequency information is often misleading, since it is possible that a crucial concept occurs only once in the document, while a non-significant one is abundant. In addition, the same concept may occur both in short and long form ('middle sized mirror with silver lining' and 'mirror').
- Concept emphasis. Emphasis arises from two sources: formatting and syntax. The first case is when concepts are included in titles and captions or have special appearances (bold, italic, underlined, etc). The second is when concepts have relevant or non-relevant position in a sentence (subject, verb, adjective and so on) or are put between parentheses. This determines not only some sort of priority, but also helps to recognize proper concept boundaries in uncertain situations.
- Concept proximity. Although closeness between concepts other than outright grammatical connection has little information for us, with careful processing it aids in identifying concept contexts or even relationship between concepts. We differentiate between concepts occurring in the same zone, those being in the same document and those connected grammatically (so a composite concept is made from them).
- Links. In World-Wide-Web documents, a particular form of emphasis constitute concepts acting as passages to other documents or document parts. Strangely, it can indicate both relevant and non-relevant concepts: the first pointing to pages containing additional knowledge, the second to ones holding more detailed information. Though a very useful notation, we will not discuss it further due to its complexity.

After extracting elementary knowledge (or more precisely, measurements) from the document, now we examine what properties and relationships can we infer from it:

- General document format and domain. If we have a dictionary listing what are the concepts typical of certain document formats (such as brochure, memo, technical data) or of given domains (for example computer science, architecture, art), comparing them with the concepts encountered in a document

roughly classify it. This needs extensive preparatory work, but in exchange it substantially improves the accuracy of further processing.

- Concept ranking. Based on frequency and emphasis, possibly taking into account relevance to the document domain discovered in the previous step, importance of a concept in the document can be estimated. Still, even if a concept is not merely a commonly used word or expression ('consequence', 'cause', 'in the light of'), mostly it only relates to the document topic, but not defines it. Another situation is when no dominant concept is found, because their ranking is too close to one another.
- Relation of a concept to other concepts. When some concepts are far more tightly connected to the given concept than others (considering proximity, frequency of co-occurrence and ranking), they are regarded as some sort of context. Context is useful not only in distinguishing different meanings of the same concept (effect of multiple senses on information retrieval is discussed in SANDERSON (1996)), but also in extending and thus specializing the concept, making it characteristic to the document topic.
- Parts of a composite concept. A special variant of the above concept relationship is between a composite concept and its parts, which is classified in four types:
  - The meaning of (supposedly primary) part and composite concept is the same, and the former is merely an abbreviation of the latter, at least inside the given document. We suspect it if the contexts of the two concepts are very similar.
  - The composite concept is a specialization of the part ('lever' and 'steel lever'), or the part is the extension in such a construct ('steel' and 'steel lever'). We consider it if the extension part is never encountered alone, but the general part is (and possibly the context of the latter contains the context of the former).
  - The (supposedly auxiliary) part serves as a distinction but only in a local context, without any general meaning ('let us assume that there are two levers, a yellow lever and a red lever'). These sorts of distinctions are usually letters, numbers or adjectives.
  - The composite concept is accidental, it does not carry any relevant meaning, only a subset of its parts do ('looking into a room full of mahogany furniture'). This situation occurs when either ranking of the concept is very low, it is too complex or only small fractions of the composite concept are encountered by themselves.

  Correct recognition of the listed relationships influences the concept structure model and may result in the re-evaluation of ranking or even omission of certain concepts. Of course, the described circumstances in which each type is likely are not exact conditions, and should be handled this way.
- Sub-topics. If the document is divided up into zones and we examine which concepts appear in each, further knowledge about concepts and their roles can be gained. We look for concepts correlated with zones, namely those

occurring in all zones and those limited to one or few – then try to find out the relationship between them. Though omnipresent concepts may be relevant as well as unimportant, the first case is more probable if the more limited concepts are specializations of the widely used ones.
- Role of a concept in the document. Obviously, this is the single most important property of a concept regarding the retrieval process, as it describes not only whether a concept is significant or not, but also answers why it is or is not. Because determining concept role heavily relies on the existence of a precise concept structure model, we postpone its discussion.

By this time it is should be clear what information should be included in the document model, which we summarize below. Do not forget that each property may be subject to modification when the concept structure model becomes known, and that in the light of these modifications, the entire document model might be re-evaluated.

- List of relevant concepts. Initially, all concepts uncovered by the syntax parser are included unless filtered according to the recognized document domain. The following properties are stored for each concept in the list:
- Rank. It specifies how much distinguishing power is attributed to the concept regarding the other documents in the collection, the greater this number, the more probable that the concept will be included in the final document representative. Instead of rank values covering a wide range, we recommend a few rank levels, since determining threshold values for future decisions becomes easier and more reliable.
- Context. Virtually a list of other concepts, for each recording some measurement of its proximity to the given concept, omitting concepts with very large distance; here, too, use of levels rather than actual numeric values is advised. Although the frequency of co-occurrence (or the rank of the relating concept) might be embedded into the distance, as being liable to possible change in subsequent iterations, we should store it separately.
- Role. Though role does not influence whether a concept will be among those constituting the representative (because rank determines that), it does specify how the concept will be employed in the retrieval process, primarily when interaction with the user takes place. We will examine role later in more detail.

5. Concept Structure Model

In the previous section, we discussed how concepts are mined and their properties determined from individual documents – a task that traditional methods carry out also more or less similarly. We could stop here and, considering the collected data as descriptive enough, begin to construct document representatives. However, these measurements can be not only improved to a great extent, but the selection of

concepts to be included in representatives also may be made more efficient regarding retrieval. Although we extracted as much information as possible about relations between concepts and a particular document, relationships between concepts in the whole document collection remain unknown.

Strictly speaking, in an ideal case connections among concepts are part of the universal human knowledge, and as such are immutable. Therefore it would seem natural that we build a large, machine readable encyclopedia defining all aspects of concept relationships (similarity, contrast, analogy and so on); this knowledge would be useful even when substituting for synonyms or conducting shallow reasoning. However, for several reasons it is unfeasible:

- Counting all possible concepts (the overwhelming majority of which would be technical terms), the number of possible two-way relations is enormous; besides, not all connections between concepts are binary – for example, analogy involves four concepts (‘leaves are the same for the branch as fingers are for the hand).
- Technical terms are not always used in the same meaning, especially in evolving domains; sometimes a well established but obsolete concept is reused in the same area. Often a new concept is introduced in a small number of articles, but never takes hold and is replaced by another. To track these volatile meanings is obviously impossible.
- Using a large static network of concepts in a document collection where document topics are focused in a relatively narrow domain frequently results in inaccurate retrieval. It happens because even a slightly altered context can mislead the discovery process into believing that the two concepts are different – as it sees several meanings.

Building concept relationship knowledge (or, in other words, describing the concept structure model) dynamically, based solely on members of the document collection, is clearly inaccurate. However, this imprecision is somewhat counterbalanced by the decision that only a limited set of relations is represented in the model and even these are handled as probabilities rather than facts. The concept structure model defines three kinds of relations between concepts:

- Specialization, if a concept is a specialization (‘bus’ and ‘vehicle’), is a part of another concept (‘keyhole’ and ‘lock’) or is understood in the domain of it (‘CPU’ and ‘computer’). The relation may be of multiple strength levels, depending on whether it represents a direct (no intermediary concept in the special-to-general chain) or indirect connection.
- Generalization – the same as specialization, but in the opposite direction. We distinguish them only for convenient reference; no method exists which would recognize this relation in one direction but not in the other.
- Correlation, when two concepts either frequently or very rarely occur together. It usually means that concepts are located in the same or rather dissimilar domains; very strong co-occurrence signifies either a more complex concept or some sort of compulsory adjective. Here, too, a few levels should

be used to qualify connection, but indication of non-correlated concepts is unnecessary.

For each concept, a general frequency index is stored, recording in how many percentage of documents the concept is encountered (similar to the inverse document frequency, see van RIJSBERGEN (1979)). Concepts with high index values are less relevant in retrieval and therefore will be excluded from representatives, for they lack differentiating power.

One sensitive issue, however, so far remained unresolved: how different meanings of a concept should be handled. A possible solution is that we distinguish literal and actual concepts, the former referring to a given sequence of letters, the latter to a particular meaning; thus relationships and general frequency indices can be separated. Still, even when we would be able to accurately determine how many different actual concepts pertain to a given literal concept, we cannot identify which actual concept is present when the literal concept is encountered in a document. Usually a simple rule of thumb is applied – if in a zone or in a document concepts seem typical to a certain meaning, then all concepts contained there will be qualified as having that meaning.

Now let us see how can concept relationships and properties listed above be recognized based on concept contexts. Of course, as document models influence the concept structure model, similarly also the concept structure model affects values stored in document models. The strength of each relation depends on three factors: rank and proximity of involved concepts, in addition to the number of cases when the given context constellation is present.

- If a concept has contexts forming a few groups, where members of a group are similar to each other while significantly differing from concepts in other groups, the concept possibly has more than one meaning. Context similarity is determined by how many percentages of concepts are common.
- If context of a concept always contains another concept, but the other concept is often encountered alone, then it is likely that the former is a specialization of the latter. Because general concepts commonly occur only a few times in the introductory sections of documents, their distance from other concepts should be decreased, so that they are not left out from contexts.
- If contexts of two concepts are similar, we may suspect that they are specializations of the same concept. Likewise, when only one concept from a certain group is present at a time in the context of a primary concept, then it is probably a generalization of all group members. Due to the initially large size of contexts, this case can be detected only in a later stage when the majority of unimportant concepts have already been discarded.
- Correlation is computed directly from concept frequency data.

When both the document and concept structure models are available, at last the concept role in the document can be determined. The following cases are distinguished along with some frequent clues for each:

- The concept is the document topic or is part of it. The concept should be among the most relevant and dominant ones in at least one document zone, meaning that all other relevant concepts must be centered on this one through specialization, generalization or co-occurrence. Role of concept in other documents does not influence our decision here, because all significant knowledge is contained in the concept structure model.
- The concept is a generalization of the document topic. Now the concept is generalization of one or more concepts included in the document topic and is encountered throughout the document in a uniform manner even if it is scattered.
- The concept is a specialization of the document topic. As follows from the previous case, the concept should be a specialization of a concept being a member of the document topic and usually occurs only in a few zones. These specializations can be mere references in the text, but if the document structure is an overview, this might explain its place in the domain, and prove valuable to an uninformed user.
- The concept is a concomitant of the document topic. Now neither generalization nor specialization, but rather the third relation, co-occurrence is present between the concept and some members of document topic. The concept is present in the majority of document zones where the correlated concept is encountered, though its rank is lower. For concepts having multiple meanings, the co-occurring concept often yields an appropriate and terse definition of a particular meaning.
- The concept is merely referenced in the document. Here the concept has a quite low rank with the only important question being whether it is related at all to the document topic or the document domain in general. If not, that may signal connection between two larger domains (helpful during user interaction at retrieval), which is then represented as a slight correlation between the two concepts describing the domains.

6. Document Retrieval

Now, as every important component became known, we are ready to discuss the entire document retrieval mechanism, depicted in the figure below. The process (see *Fig. 1*) consists of three stages: the query-independent off-line stage, where the initial document models are built; the refinement stage, when document and concept structure models are synchronized; and finally the query-dependent on-line stage, where the user submitted queries are answered in an interactive fashion (described later) and user behavior is employed to improve the accuracy of document models.

The off-line stage includes the following pre-processing steps before or during the shallow syntax parsing applied to mine concepts:

- Removing special punctuation, such as quotation, dialogs, exclamations and others in order to simplify the task of syntax parsing; besides, incomplete

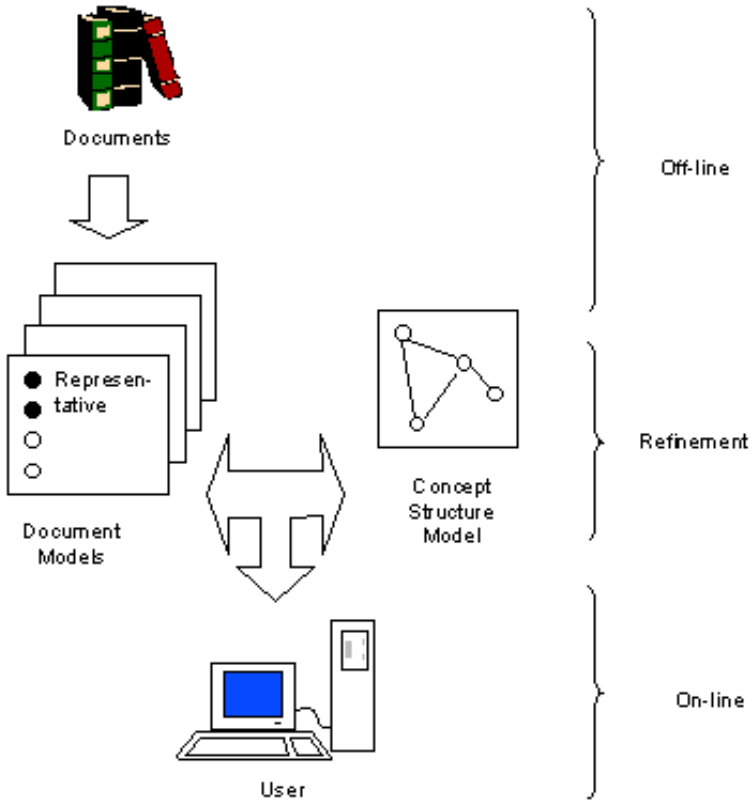

Fig. 1. Overall architecture

and short sentences may be omitted. Taking into account stylistic and modal information imposes an unjustified computational burden.

- Recognizing and replacing synonyms with a uniform word or expression; better yet, if we can translate expressions into canonical codes looked up in a dictionary – then translation between different languages becomes possible. However, since meaning is very context-dependent, this processing can be carried out only partially.
- Simplifying grammatical constructs which includes word re-order and splitting complex sentences into many simpler ones repeating certain sentence elements in each, if necessary.
- Replacing pronouns with the corresponding noun or proper noun construct. Although in some languages this task needs a sophisticated algorithm (and ambiguous cases should be left unresolved), it is utterly important to discover

as much occurrence of a concept as possible, as the more context data available, the more precise the different models, improving retrieval performance.

The refinement stage may consists of multiple iterations, where at each iteration information gained from document models are applied onto the concept structure model, which in turn affects values in document models; processing ends when no relevant change is made in either model. The stage is invoked at the initialization of the retrieval process itself, whenever a document is added to or removed from the document collection, and when a user interaction occurs. Refinement can mean the following:

- Model value modifications. It occurs, for example, when a new relation is established among two concepts, or when rank of a concept is increased. Causes and effects are presented in detail in the next section.
- Concept removal. If a concept is detected as ubiquitous or, on the contrary, as too scarce and unrelated to any other concept, discarding it simplifies and accelerates further processing. Still, the concept list in a document model is not limited to concepts included in the document representative, because future refinements may advance concepts into or withdraw them from it. Therefore, particular caution is needed as to which concepts are lost for good.
- Concept addition. Recognition of the document domain or format means addition of the suitable concept(s) to the document model. In addition, it is often useful to include generalized (up to that describing the whole document collection, if it is found) or co-occurring concepts of representative member concepts (as in query term extension, see van RIJSBERGEN (1997)). However, they should be marked so that when rank of the primary concept decreases placing it outside of the representative, these complementary concepts are canceled.
- Concept generalization. There are some circumstances when replacing many specialized concepts with one concept being a generalization of all of them improves performance, as more context information can be extracted. We regard it as an alternate form of synonym replacement.

The on-line stage is executed each time a user submits a query. First, the query is processed in a fashion similar to document pre-processing – but now only synonym replacement is employed and words not grammatically connected to the core primary concept are discarded. This concept, called the initial concept, is the starting point of an interactive session, where any additional query refinement is made through a series of selections from presented options. (From the point of view of user, he or she simply browses some kind of a concept hierarchy.) Conducting the session this way has two purposes: on one hand the user always has an overview of all possibilities, even when the domain is unknown for him or her, and on the other hand, the retrieval process sees user behavior in a larger context, hence exploiting it much more efficiently.

At each point in the on-line stage there is a so-called focus concept (at the beginning the initial concept), around which the currently displayed document references and concepts are organized. Part of a sample query result page is shown in *Fig. 2*. The following lists are presented for the user (each list member can be selected for further examination, meaning either a document or a renewed focus concept):

- When the focus concept is a single word having multiple meanings or an expression which might be interpreted in different ways, the result (including every list enumerated below) is split accordingly.
- If the focus concept is fairly narrow, list of documents whose topics (more precisely, representatives – relevant part of document models) are very similar to or the same as the focus concept; otherwise list of concepts which occur in document topics beside the focus concept. When some classification exists in the documents (for instance author, department, location), grouping according to them should be performed.
- List of concepts which are different level generalizations of the focus concept; the user can extend his or her search criteria selecting these concepts if the result contains no or only few references.
- List of concepts being immediate specializations of the focus concept. Although with the tools at our disposal it is impossible to precisely build the concept structure model and, naturally, document collections do not cover completely every nook of their domain, determining which concepts are 'sisters' in the hierarchy is utterly important. Not only because the amount of all specializations can be overwhelming, but also to fully exploit advantages of the browsing approach.
- List of concepts co-occurring with the focus concept. When among the documents containing the focus concept in their topics some are inhomogeneous (discussing more concepts with an equal emphasis – comparisons, evaluations and so on), co-concepts should be included in the list. This way connection between domains can be comprehended.

User behavior materializes itself merely in selections performed during a session, since asking users to qualify results obviously works solely in an experimental environment, where only a small amount of data is gathered. However, depending on where the given selection is made, it is interpreted differently: choosing a co-occurring concept strengthens correlation between the concept and focus concept, while selecting a document, on the other hand, decreases slightly the rank of concepts forming topics of the other documents. We should take user actions into consideration rather carefully, as they may originate in interest, ignorance or curiosity.

Focus: land

Meaning #1:

Documents –
Classification of agricultural areas (#9821)
Land usage characteristics (#1293)
Color patterns in satellite images (#3457)
Seasonal changes and observation (#3445)

Generalizations:
Area, Object

Specializations:
Residential area, Highway, Meadow, Corn field . . .

Related:
Satellite, Image, Observation, Usage, Colour . . .

Meaning #2:

Documents –
Automatic landing procedures (#2345)
Special-use aircrafts (#3439)
. . .

Fig. 2. A sample query result page

7. Algorithm

Instead of describing algorithms for document and concept structure model building and refinement, we give a more or less exact enumeration of where the data might come from for making decisions, though what these decisions should be remains an area of future experimentation. First we examine what relationships can be detected regarding concepts; second, connection among the two models and user behavior is presented.

As mentioned in previous sections, analyzing documents results in frequency and co-occurrence data for each encountered concept. We inspect relations always between a single concept and other concepts, using several aspects:

- Analysis according to scope. We examine with which concepts a given concept often or rarely occurs in the same grammatical structure, document zone, document or document collection (in other words, in contexts of various strength). It should be remarked that correlation does not mean simply

'together' and 'not together' measures, also more complex patterns exist, for instance 'if this concept is present, the other one occurs too, but not the other way around'. In addition, not only binary correlation can be looked for – for example 'concept A is encountered either with concept B or concept C, but never with both'. Comparing the list of correlated concepts at each scope level and at each element at the same scope level yields valuable information.

- Comparison of concept contexts. Given two concepts, we analyze that at different scope levels (as mentioned above) and at various elements at the same scope level how much their contexts agree or differ. This process involves far more comparisons (growing not linearly but rather exponentially with the number of concepts) than the previous one, so preliminary concept elimination is inevitable. Even complying with that, ternary or quaternary examination remains out of reach.
- Analysis of concept role. The most intricate process of all, here we compare how a given concept is qualified (relying on observations from the above procedures) in contexts of concepts which are members of the context of the given concept. Usually only the strongest context (concepts in the same document zone or being high ranked) is worth examining, and even this for solely the most relevant concepts. Since results may be very diverse, a sophisticated evaluation is required.

Connection between document models and the concept structure models is two directional, as implied from the way synchronization between them is performed, as described in the previous section. Let us consider first how document models affect the concept structure:

- Rank. When any kind of relationship is present between two high ranked concepts, because they are qualified as discussed in detail in the given document, their connection in the concept structure model also should be made stronger than relations between lower ranking concepts.
- Context. Pattern of context variation at various scope levels and in different elements (zones or documents) determines the sort of relationship, which should be built in the concept structure model between the two concepts.
- Role. Though role is document centric, it is only indirectly based on observations made while constructing the document model, rather it originates from discovered knowledge in the concept structure model. An advantage of this relative independence is that value fluctuations during refinement iteration cycles affect role to a lesser degree.

Influence in the opposite direction is quite restricted due to the fact that the document model is and should be closely related to the structure of individual documents. Concept rank is slightly, concept role is more intensively linked to the concept structure model as follows:

- Generalization and specialization. If in the context of a concept we can find concepts being specializations or generalizations of it, that means an extensive

and justified presence in the document, therefore an increased rank. Concept role is established by looking at how relevant concepts are connected to each other in terms to special-to-general.

- Correlation. When co-occurrence of two concepts is detected, usually one of them is less important than the other and thus its presence in the document model does not carry significant information. Similarly, if two concepts do not occur together, then in the hence rare case when one is encountered in the context of the other, its rank should be lowered. The same is true for concept role.

- General frequency index. Unfortunately, frequent concepts may be as important or non-relevant as scarce concepts; however, when we regard it along with a high ranked concept, rarity means a possible distinctive power, and as such it implies an increased rank.

Influence of user actions to both models is rather straightforward, and does not need an elaborate explanation. However, it is important to notice that since both selection and non-selection matters, a single user action will not cause any model value modification. Moreover, as user actions in lists containing different members cannot be compared, an enormous amount of recorded usage information is needed to securely cover the majority of possible cases. Selection in lists describing other concepts relating to the focus concept increases, while non-selection decreases the respective type of relationship between the two concepts. On the other hand, when the list comprises document references, selection increases and non-selection decreases ranks of all concepts included in the representative of the corresponding document.

Fig. 3 shows the parameters which should be taken into account during an implementation effort (a plus sign means that the given feature is recommended, while a minus sign means that it should be avoided).

As it can be seen from *Fig. 3*, the most important goal in our opinion is to improve the document retrieval accuracy (precision and recall), even if it requires heavy computation during the off-line processing stage. Owing to the comprehensive nature of the statistical analysis, documents cannot be processed completely serially in this stage. This means that a large amount of descriptive information relating to documents should be kept in memory, suggesting a distributed computing approach – however, the memory need can be alleviated somewhat by aggressively reducing the number of considered concepts before the concept structure model is built. Another possible solution is to start from a few well chosen documents, then successively refine the concept structure model by including more and more documents (a particular way of synchronizing the document and concept structure models). Although this method requires less memory, the cost of multiple iterations can easily diminish that advantage.

Feature	Advantage	Drawback
previous knowledge about document formats, domain vocabulary, concept structure (+)	more accurate document model	must be maintained and updated frequently (1)
deep syntax analysis (–)	more accurate document model	high computational cost (2)
deep statistical analysis (+)	more accurate concept structure model	very high computational cost, possibly contradictory results
successive refinement of the document models and the concept structure model (–)	more accurate retrieval	risk of an instable concept structure model
heavy use of calculated concepts (3) in representatives (–)	reduced storage size for document representatives	the differentiating power of representatives deteriorates
concepts have wide context in the document text (+)	more stable concept recognition	statistical analysis requires more resources

Notes: (1) automatic thesaurus construction is an example for maintaining domain specific and lexical knowledge without human intervention

(2) experience in the area of information extraction shows that syntactic and semantic analysis does not enhance precision significantly (see GAIZAUSKAS and WILKS (1998))

(3) generalized and co-occurring concepts are added to the core concepts during retrieval, based on the concept structure model

Fig. 3. Implementation considerations

8. Future Plans

The retrieval system outlined in the present article is not implemented yet, so no test results and therefore no comparisons are available with existing systems; and without these data, the value of our contribution remains unknown. Consequently, implementation is the primary concern, during which we seek answers to the following questions:

- Which statistical measurements are significant when determining the concept structure model?
- Regarding the concept structure model, what is the upper limit of accuracy attainable with statistical analysis?
- How much can be gained when processing the document collection as a whole as opposed to individual document processing?
- What techniques can reduce the heavy computational requirements our method entails?

A further point of interest is to examine how the system behaves when members are removed from or added to the document collection, or when their content

is modified – how the quality of the concept structure model deteriorates and what can be done to prevent it.

References

[1] BEAR, J. – ISRAEL, D. – PETIT, J. – MARTIN, M., Using Information Extraction to Improve Document Retrieval. *Proceedings of the Sixth Text Retrieval Conference*, Gathiersburg (MD), November, 1997, pp. 367–377.

[2] COLE, R. – EKLUND, P., Application of Formal Concept Analysis to Information Retrieval using a Hierarchically Structured Thesaurus. *Proceedings of the Fourth International Conference on Conceptual Structures*, Sydney, Australia, 1996, pp. 1–12.

[3] COHEN, W. – SINGER, Y., Context-sensitive Learning Methods for Text Categorization. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1996. ACM Press, New York. pp. 307–315.

[4] EVANS, D. – ZHAI, C., Noun-Phrase Analysis in Unrestricted Text for Information Retrieval. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, California, June 1996, pp. 17–24.

[5] FALOUTSOS, CH. – OARD, W., A Survey of Information Retrieval and Filtering Methods. Technical Report CS-TR-3514, Department of Computer Science, University of Maryland, 1995.

[6] GAIZAUSKAS, R. – WILKS, Y., Information Extraction: Beyond Document Retrieval. *Computational Linguistics and Chinese Language Processing*, **3**, No. 2, August 1998, pp. 17–60.

[7] LALMAS, M. – RUTHVEN, I., A Model for Structured Document Retrieval: Empirical Investigations. *Hypermedia – Information Retrieval – Multimedia*, September 1997. Dortmund, Germany. pp. 53–66.

[8] van RIJSBERGEN, C. J., *Information Retrieval*. Butterworths, London, England. 2nd Edition.

[9] SANDERSON, M., Word Sense Disambiguation and Information Retrieval. Ph.D. Thesis, Technical Report (TR-1997-7) of the Department of Computing Science, University of Glasgow.