Principles of content analysis for information retrieval systems: an overview

Krause, Jürgen

Veröffentlichungsversion / Published Version
Konferenzbeitrag / conference paper

PRINCIPLES OF CONTENT ANALYSIS FOR INFORMATION RETRIEVAL SYSTEMS: AN OVERVIEW

JÜRGEN KRAUSE

Unquestionably, the content analysis which has emerged as part of Information Retrieval Systems (IRS, e.g. literature databases) over the past 20 years has much in common with the content analysis used by linguists or in the social sciences. However, its intrinsic value stems from the special context in which it is used:

a) Close interdependencies link the selected content analysis with the retrieval situation. The user's retrieval strategies, which are intended to obtain information relevant to the current problem situation, and the available aids (e.g. expansion lists or user-friendly browsing tools) affect the efficacy of some analysis techniques (e.g. noun phrase analysis from computer linguistics) to a considerable extent.

b) Normally, a commercial IRS handles mass data, thus necessitating the use of a reduced content analysis even today. Full morphological, syntactic, semantic and pragmatic text analyses are unthinkable simply for efficiency reasons but also for knowledge reasons. Content analysis in IRS is therefore a component part of a special type of restricted system which obeys its own laws.

Against the backdrop of these considerations, forms of content analysis in present-day commercial retrieval systems are studied and promising expansions and alternatives are proposed.

1. Introduction

The objective is to show possible approaches for improving retrieval functions of IRS based on the state of the art now attained in commercial systems and practice-oriented

developments in research, and to determine the advantages and disadvantages of individual solutions. Since content analysis measures can be frequently exchanged with those on the retrieval side, both the aspect of content analysis and retrieval will be examined. A certain type of content analysis may therefore only be chosen to organize the retrieval algorithm efficiently. A simple example is the retrieval function of truncation. It is largely superfluous if compound splitting and basic form reduction are used in content analysis. However, compound splitting and basic form reduction can also be replaced by equivalent generation methods during research. In an ideal situation, the user does not notice whether an algorithm expands the user's search word to include all word forms or whether the word forms of the document are reduced to basic forms when descriptors are allocated.

Commercial text IRS are now primarily based on intellectually or automatically determined descriptors which are researched with or without additional thesaurus relations by means of Boolean algebra. The following comments are restricted to this type of research and the overcoming of its inherent weaknesses by computer-linguistic and quantitative-statistical methods. In addition, information science is aware of other basic types of retrieval which each attract other or modified content analysis methods. The best-known and most widespread additional types of search are:

a) Search path organized on a hierarchical basis

Due to an ever greater restriction on selected generic terms, the user is guided to his target data by means of a path specified by the system. All selection alternatives are provided on the user interface. Active noting features are not required. The user need not search for his own terms, instead he (she) makes a selection from a displayed repertoire.

However, these advantages are accompanied by marked disadvantages: the low quantitative limits of this solution must be taken seriously. If the hierarchical structure (especially depth) increases, clarity on screen is quickly lost. It depends a great deal on the (normally objective) conditions of the area of application whether a hierarchical system can be considered. Hierarchical access alone is only adequate very seldom.

In the case of hierarchical systems, the advantage of content analysis primarily lies in the intellectual construction of the flow chart and allocation of the individual documents to the nodes.

b) Hypertext relations

Kuhlen (1991) extensively examined possibilities and limits of hypertext systems and their realization forms. Hypertext realization should be regarded as an additive element

of conventional descriptor systems in the same way as hierarchical access - supplementary to special search situations - can improve descriptor systems.

In addition to a) and b), the information science discussion is aware of a number of other types of research. In Bates (1989:412), for example, the traditional descriptor system is only one of seven basic types which would all have to be supported in an ideal text research system:

> "1. *Footnote chasing* (or 'backward chaining' ...). This technique involves following up footnotes found in books and articles of interest, and therefore moving backward in successive leaps through reference lists...
>
> 2. *Citation searching* (or 'forward chaining'...). One begins with a citation, finds out who cites it by looking it up in a citation index, and thus leaps forward.
>
> 3. *Journal run*. Once, by whatever means, one identifies a central journal in an area, one then locates the run of volumes of the journal and searches straight through relevant volume years.
>
> 4. *Area scanning.* Browsing the materials that are physically collated with materials located earlier in a search is a widely used and effective technique. Studies dating all the way back to the 1940s confirm the popularity of the technique in catalog use.
>
> 5. *Subject searchers in bibliographies and abstracting and indexing (A & I) service*. Many bibliographies and most A & I services are arranged by subject. Both classified arrangements and subject indexes are popular. These forms of subject description (classifications and indexing languages) constitute the most common forms of 'document representation' that are familiar from the classic model of information retrieval discussed earlier.
>
> 6. *Author searching*. We customarily think of searching by authors as an approach that contrasts with searching by subject. In the literature of catalog use research, 'know-item' searches are frequently contrasted with 'subject' searches, for example. But author searching can be an effective part of subject searching as well, when a searcher uses an author name to see if the author has done any other work on the same topic."

The expansion of the possible types of research in an IRS is an important subject for the further development of current commercial approaches. Due to reasons of space, this subject must unfortunately be omitted along with components of intelligent Information Retrieval (IR) which were already discussed in Krause (1992).

2. Principles of content analysis and information retrieval

By way of introduction, it is necessary to clarify some fundamental principles and problems of IR which are all ultimately rooted in an indistinct way at all system levels (Krause, 1990). This generally occurs - although content analysis is a main subject - from the viewpoint of search, as every problem appears to the user as a search and interaction problem. He (she) also experiences all content analysis measures as indirect impacts on his (her) search formulation, which is why conceptualization of the entire process should be seen from this aspect.

2.1 Descriptor search using Boolean algebra for document retrieval

Boolean algebra can be used to establish a link between queried terms by means of the logical operators AND, OR, NOT. These are often supplemented by formal additional techniques such as truncation (right, left, inward truncation) or neighborhood search (for closer definition of the AND operator) which function exclusively through exact-pattern-match processes.

During document retrieval with Boolean algebra, the user can normally fall back upon two types of descriptor regarding a document:

a) Aspect-related descriptors such as name (obligatory), organizational code, author and (modification) date and

b) Free descriptors (= keywords) which can also be obtained from the already existing stock of descriptors by marking the list entries (= connection with the model character).

From the viewpoint of content analysis (indexing), the descriptors - irrespective of any syntactic or hierarchical reference (more generally: without any relationing with each other with the exception of some aspect statements) - will characterize the contents of the document. Users utilize the same unrelationed, content-identifying terms in the search, thus avoiding the mentioned difficulties with hierarchical access.

2.1.1 Thesauri

Commercial descriptor systems normally permit a number of general connections between individual terms (synonyms, associations, generic terms, subterms, etc.). Thanks to these global relations, a list of descriptors turns into a thesaurus. If the thesaurus contains generic terms and subterms, an attempt is made in a descriptor system containing no

second search path to integrate the advantages of an hierarchical system in the list of descriptors (cf. DIN 1463).

Nowadays, there is no longer any doubt concerning the general effectiveness of thesauri which are generated intellectually or semi-automatically to improve the widest range of IRS (see, for example, the Darmstadt indexing approach: Lustig, 1986; Fuhr, 1988; Biebricher et al., 1986).

2.1.2 Intellectual versus automatic descriptor determination

The discussion concerning both these basic types is as old as the first IRS used in practice. For example, the Informationszentrum Sozialwissenschaften (IZ Bonn, 1994), develops its literature databases fully intellectually on the basis of a thesaurus which is organized partially according to a hierarchy. The indexer takes all allocated terms (exception: additional field for "free" terms) from a thesaurus which must be constantly updated. The terms contained in the thesaurus form an integrated semantic system. The advantage of these terms is that the indexing depth can be controlled during generation of the thesaurus and semantic standardization is enforced at the level of indexing. However, the terms in the restricted, specified vocabulary can "lose" their colloquial semantics and can be almost interpreted in formal language. This characteristic of controlled, intellectual indexing becomes most apparent when an indexer does not find a desired term in the thesaurus.

Purely intellectual indexing based on controlled thesauri produces excellent results in some areas. However, this approach is hardly used any more in practice for large data stocks since the costs of intellectual processing are regarded as too high. Users of these systems also do not normally consult the list of keywords, but formulate their query directly. If the thesaurus is not constantly updated to include new (fashionable) terms in a specialist area, this results in excessive discrepancies between the terms selected by the users and the thesaurus structures.

Moreover, it has not yet been possible to the best of my knowledge to prove the postulated advantage of intellectual analysis using a controlled thesaurus by means of corresponding evaluation.

Intellectual processing has largely been replaced in commercial systems by automatic free-text methods in which thesauri and, if necessary, computer-linguistic methods (e.g. basic form tracing) are only still used in sub-areas to resolve the linguistic diversity of the starting texts.

The DATEV databases are, for example, an extreme example of the "pure form" of traditional, automatic free-text systems (cf. DATEV, 1994). Except for some aspect-related descriptors, only a stop word list regulates the selection of descriptors. The JURIS databases (GOLEM/PASSAT base, cf. Möller, 1993) are another example. It is interesting to note here that after more than fifteen years practical experience, intellectual analysis, which was vehemently rejected at the start of development, is again regarded as a cure for the empirically observed, poor retrieval performances of JURIS (cf. Wolf, 1992; Möller, 1993).

2.1.3 Summary

With all of the above-mentioned methods, users often gather during their work extensive heuristic knowledge on what descriptors can be formulated adequately or as search queries. Users also become specialists regarding Boolean retrieval. The basic principle of the inverted list also produces good response times because the actual data stock need only be accessed when the user wants to examine document texts.

However, these advantages of conventional Boolean retrieval are contrasted by marked disadvantages.

Before they are discussed in detail, the theoretical basis of the so-called standard model will be explained more fully.

2.2 Standard model of traditional information retrieval

This is also called the Salton paradigm (cf. Belkin, 1993:57). Basically speaking, it remained the starting point of any suggested improvement regarding IR up to the 1980s. The central element in this model is the exact pattern match method which combines the terms selected by the user with those in content analysis. As shown above, only the logic operators AND, OR, NOT, brackets and formal additional techniques such as truncation or neighborhood search are available (e.g. restricting AND to the sentence through the corresponding context operator).

The methods described in the next sections improve the pattern match method by eliminating some obvious weak points (in particular, section 3) or replacing Boolean algebra by quantitative-statistical procedures (cf. section 4), but do not affect the general validity of the basic model.

Its main weaknesses become most apparent when the user modifies the query after the system has supplied the initial results:

> "Concern with representation of the information need has typically arisen after the process of judgement, which is typically to be performed by the user, as an estimate of the potential relevance of the text to the information need. The results of the judgement process are then used by the system to modify the query, or, occasionally to modify the text representations. This process of... 'relevance feedback', is perceived ... as an attempt to gain the 'best' possible representation of the user's query ... that is, to improve the representation so that the comparison process will work most effectively. It is important to recognize that, in this model of IR, the person involved in IR is seen as a user of the system, standing outside of it. Involvement of the user with the IR system is minimal and interaction (in the form of the judgement process) is seen as ancillary to, and only in support of, the representation and comparison processes." (Belkin, 1993:56).
>
> "The force of these assumptions is ... to devalue or even ignore the significance of interaction of the 'user' with the texts; and to provide support for only one form of information seeking behaviour, that associated with searching for some well-specified item. Additionally, through the privileged position of comparison and representation, and the assignment of responsibility for these activities to the system, the standard view of IR leads to strong control by the system of the entire IR process, and the consequent lack of power or control by the user" (Belkin, 1993:57).

In the discussion of the principles examined here, this global criticism is used more as a general guideline to remedy the weaknesses of the standard mode by additional measures. Only the discussion - mainly omitted for reasons of space - concerning components of an IR and that concerning additional search types force us finally to leave the model (cf. section 1).

3. Computer-linguistic methods as a supplement to free-text retrieval

Improperly from the viewpoint of computer linguistics, traditional free-text retrieval reduces content analysis to a symbol-oriented analysis (cf. Knorz, 1994). Documents are regarded as the stringing together of chains of symbols which are separated by blanks or punctuation marks. Every morphological, syntactic or semantic item of information contained in the text is classified as negligible, which is the reason why it is not analyzed.

The designers of these systems (hereinafter described by means of examples in German) naturally know that the omitted morphology, syntax and semantics shorten content

analysis, i.e. for the most part irreversibly: syntactic information in the documents is no longer available - once erased during document recording - when the query is made (cf. the classic example: *Suche nach griechischen Schiffen, die römische Häfen anlaufen* (look for Greek ships which call at Roman ports*)* versus *römische Schiffe, die griechische Häfen besuchen* (Roman ships which visit Greek ports). The methods developed for this purpose are not dependent on language and can therefore be used internationally. Due to their simplicity, they can be developed in an extremely robust manner for mass data. They are also fully automatic, quick and cheap. Symbol-oriented, general methods on the query side will counter the lack of analysis depth at least in the area of morphology and, at times, in syntax. Truncation and context operators are actually operations without "linguistics", but the developers of the IRS are confident that the user has this knowledge (e.g. *Haus** for all compound words such as *Haustür, Hausverwaltung,* etc.) and uses it adequately in search strategies. By resorting to user knowledge, retrieval-side truncation will therefore solve the problem of word form combination, compound splitting and derivation combination, and intelligent use of context operators will replace the lost references of sentence structure.

3.1 Methods available

The computer-linguistic argument assumes that the reduction of content analysis to a purely symbol-oriented, non-language-dependent dimension is responsible for the majority of research problems. The language-dependent rule systems researched by linguists are therefore integrated. They replace the user's linguistic skills which he (she) must transfer during traditional free-text research to the symbol-oriented help operations (truncation, context operators) which are by no means suitable for this purpose. The quality of the exact pattern match method of traditional IR systems will be improved by:

a) returning the word forms in the text to their basic forms or expanding the word forms of the user query to all related basic forms. The recognition of abbreviations can be classified here as a special case.

e.g. *Stall: Stalls, Ställe, Ställen - Thema: Themen, Themas, Themata*

b) Splitting compound words into their constituent parts

e.g. *Druckerzeugnis: Druck-Erzeugnis* versus *Drucker-Zeugnis*

c) Combination of derived terms

e.g. ADJ *lieblos* / NOUN *Lieblosigkeit* (*-keit* turns adjectives into nouns);

Formatierung / Format / formatieren

Rang er nach Luft: basic form *Rang* or *ringen*

d) If the structure is given related terms (multi-word terms, complex descriptors, verb prefixes such as *gehört ... zusammen*)

e.g. *fing ... an ...*: basic form *fangen* or *anfangen - hielt ... den Atem an*

• natürliche und juristische Personen - kalter Kaffee (= Spezi) versus kalter und abgestandener Kaffee

e) Hyphenated part-word deletions e.g. *Haus- und Hofwirtschaft*

f) Phonological-graphematic translation of proper names and spelling checker

Spelling checkers are a standard feature of text-processing programs and are supplied by special companies for a large number of languages. Curiously, most information systems do not have a spelling correction feature for the search term. It should be as natural as a check on the entered texts. This is especially important if computer-linguistic methods supplement free-text search.

A special case in spell checking is verification of names: *Maier* instead of *Meier* is not entered by the user within the meaning of a spelling mistake; he (she) does not know the correct spelling. In many cases, pronunciation is the link between the different spelling variants. The user often only remembers the sound of the name. It is therefore necessary to provide an additional computer-linguistic function which first transforms the entered term into a phonetic transcription from which all possible graphematic terms are generated (see, for example, Regensburg Phonology: Hitzenberger, 1987).

g) Standardization of spelling

Especially in the case of German texts, databases often contain standard deviations from Duden spelling. These deviations lead to errors if the user is utilizing the "correct" spelling during his (her) search.

• Umlauts or ablauts replaced by vowel + e: *mueßig. müßig*

• Capitalization of words written with a small letter at the start of the sentence

• *ß* stored as *ss*

Capitals as highlight in the text and mixed spelling:

DATEV: Datev; GmbH: GMBH

• Special without blank character: *Herr/Frau* instead of *Herr / Frau*

In the near future, attention will also have to be paid to the implementation of the moderate spelling reform of German which will probably be adopted in 1995.

h) Stop word lists / negative lists

Traditional stop words are all functional words in a language (a*nd, the, a, if,* etc.) which themselves do not have any semantic meaning, but define the relationships between terms. During automatic free-text retrieval excluding computer-linguistic methods, they therefore become "meaningless" because the relation system in the text is deliberately regarded and deleted as irrelevant information (cf. Ruge, 1994b).

Stop word lists normally reduce the descriptor list by around 50%. However, they do not affect the quality of research since users do not select them as search terms, only as processing advantages. A problem occurs in that they often do not contain uncovered homonyms (e.g. *Nur die Firma NUR).*

Additional computer-linguistic algorithms will mainly increase recall, but the different types of word combinations will improve precision.

In the simplest (and most frequently implemented) case, computer-linguistic methods work via an isolated symbol chain. This is generally possible in all cases except during word combination and with hyphenated words. However, the reduction in basic forms can be improved by means of context-sensitive heuristics (partial parsing).

- Syntactic level: *Ehe er ....*: the noun *Ehe* is not a stop word as *er* follows.
- Statistical probability: *Trotz vieler Vorkommnisse*: *vieler* is only used very rarely after the noun *Trotz*
- Semantic homography: *Das Geld auf der Bank*: *Geld* excludes the meaning of *Bank as somewhere to sit*.

Multi-word groups are generated in some systems (see, for example, the Saarbrücken system CTX, Schneider, 1987) by means of complete syntactic analyses which can also be interpreted as precise linguistic definition of context operators (in the same nominal phrase, in the same sentence, etc.).

The limit of traditional free-text systems is attained when the algorithms require a generally different type of contents representation. For example, CONTEXT (GMD IPSI, cf. Haenelt, 1994) analyzes the text of the document along syntactic, semantic and pragmatic lines, stores the thus retained complex contents structure as a conceptual network and must analyze the user query using the same method.

The RELATIO/IR system by IBM uses an interesting temporary form (Rahmstorf, 1994). In this system, the thesaurus itself is organized as a controlled, semantic network. The semantic analysis of nominal phrases in the document reorganizes the user interface structure of the search query in a subtree of the thesaurus network. Since all terms are derived from the thesaurus and are therefore used as "controlled vocabulary" in the more

complex form of the thesaurus, new contents cannot be indexed before the thesaurus is extended intellectually.

Summary: there are a whole range of computer-linguistic modules which can be used to improve content analysis of IRS further. Their use in commercial systems now no longer poses a problem for more complex languages such as German.

Efficient algorithms with satisfactory performances are, for example, offered by SOFTEX (cf. Zimmermann, 1993) or LINGSOFT (GERTWOL product; cf. LINGSOFT, 1994). GERTWOL was the winner of the 1994 MORPHOLYMPICS, a competition organized by the GLDV, in which the computer-linguistic methods being discussed here were subjected to an (informal) comparative test (cf. the special publication of the LDV forum in June 1994).

3.2 Effectiveness of computer-linguistic sub-algorithms

After discussing the general options of integrating computer-linguistic methods, the question remains as to whether the additional time and effort are really worthwhile. If this question is examined, it is noticeable in particular that very clear opinions are normally expressed here. In fact, almost everyone knows in advance what is produced by a retrieval test. Unfortunately, everyone is convinced of the superiority of another system.

This fact gives rise to spurious plausibilities with high superficial persuasiveness, which largely prevent a more in-depth examination of the effectiveness of computer-linguistic components.

3.2.1 Graduated model: additive supplements

The arguments used in research literature and among practicians in favor of computer-linguistic algorithms can best be illustrated by means of a graduated model of linguistic components:

a) Objections are raised against the pure free-text solution because the formal options offered in the retrieval language (for example, truncation and context operators) are insufficient to allow error-free recording of the various word forms of a term. The possibility cannot be ruled out that the user will not think about individual word forms or fully understand the side-effects of truncation (ballast). The attention of the user will also be diverted. He (she) must deal with purely formal considerations, a situation which irritates him (her) in his (her) task of finding the correct content descriptors.

b) In the case of algorithms which are limited to morphological analysis and compound splitting, objections are raised to the effect that they are inadequate. The content terms are related structurally in the document. This becomes most apparent with nominal phrases. For example, a user looking for the complex descriptor *Aufnahmevorrichtung für Kernspinresonanzspektren* does not actually examine documents relating to "Aufnahmevorrichtungen" or which have any connection with "Kernspinresonanz", but documents which connect both terms with the factor *für*, i.e. contains them in a special relation in terms of contents.

c) Objections are raised against syntactically-oriented methods because they only depict structural regularities but do not always agree with content relations. For example, studies of patents in the PADOK-I project (cf. Krause, 1987, Womser-Hacker, 1989) showed that the syntax analysis of the Saarbrücken system CTX only determined 75% of the text relations which agree in terms of contents with the complex descriptor of the search query (cf. Schneider, 1987).

d) Kuhlen argues as follows against the restriction to a)-c):

"The contribution of morphological, syntactic and semantic analysis to the overall performance - "Provision of information" - may be minute. As a result, it is not worthwhile to replace the dominant ad hoc methods - e.g. [...] context operators instead of syntactic parsings [...] by linguistically justified methods." (Kuhlen, 1985:7).

It also corresponds to the general pattern of these arguments that the actual increase in value will result from the component to be supplemented. The content analysis functions of each lower level are classified as not effective enough.

If the linguistic subcomponents of the graduated model (morphology, syntax, semantics and pragmatics) are analyzed as a whole from which no component emerges, there is no sensible reason to agree with the idea behind this chain of arguments, i.e. the thesis that a complete morphological, syntactic, semantic and pragmatic analysis promises the best retrieval results in terms of quality.

Nothing appears intuitively more plausible than to demand that all processes which can be observed in data processing and generation by humans should be automated as far as possible and in a 1:1 relation. At the same time, this is the simplest and most problem-free way to determine the basic design of an information system. There is nothing more undemanding from conceptualization and nothing more complicated than to take individual components from this so understood simulation approach and to justify their selection stringently.

Consequently, there is also no reason not to proceed in this way unless - seen in quite general terms - you do not *want* to (e.g. faced with a 'morally' judging background as in Weizenbaum, 1976) or *cannot*. Both cases mean that - for whatever reason - parts of the individual components shown in the graduated model no longer apply. However, if the complete chain of individual components is not implemented in full, 1:1 simulation of human data processing and generation switches to the type of restricted systems. We are therefore faced with the question of what components these restricted systems should contain and how such a choice can be determined and justified.

With regard to current commercial IRS, it is relatively easy to answer the question of whether all information-linguistic components of the graduated model can be realized. No fully developed systems in a commercial sense are available for the components of semantics and pragmatics. Even experimental systems such as CONTEXT or RELATIO/R (cf. section 3) only cover part of the semantic references contained in a text. Content analysis must therefore by necessity be designed as a restricted system.

The following simple example will illustrate that more computer-linguistic algorithms can not only have no effect in this context, but can also cause further damage (cf. Krause & Womser-Hacker, 1990 for further results of empirical studies).

During the PADOK-I project, the following sequence came about during extensive research tests involving data from the German Patent Office (text basis of the documents: title + abstract, cf. Krause, 1987, Womser-Hacker, 1989) relating to recall (percentage of relevant documents found):

a) Free text never attained the top position although it was clearly favored in the test design.

b) In the overall evaluation, PASSAT (for example, morphology, compound splitting) led to better recall compared with CTX (additionally complex descriptors from a noun phrase analysis). This result was statistically significant in a user group (industry/technical information centers). However, better precision (percentage of ballast) was achieved during retrieval using the database connected to CTX.

An industrial user made the following comments based on his experience with CTX and PASSAT:

"When working with the CTX database, I was irritated by having to think about the complex descriptors during formulation of my own retrieval strategies. The CTX component of the complex descriptors, behind which the additional syntax analysis of CTX stands, meant that I had to consider constantly which text passages were recorded and which were not by the complex descriptors. With PASSAT, however, the basic effect

could be followed relatively easily and incorporated in the cognitive process of formulating a retrieval strategy."

The comments of the industrial user therefore refer to a side-effect of the use of CTX and to additional cognitive stress which might interfere with a potentially positive effect. It is important that this side-effect is not produced through the operation of functionality in the user interface, but is based on the supplied functionality itself. It is not the actual functionality that produces the result, but that what is triggered by this functionality in an overall context.

Consequently, it must always be expected that computer-linguistic components induce processes whose impacts can no longer be controlled analytically (see Krause & Womser-Hacker (1990) for other, even more complex examples). The question of the sequence of systems with different strengths of computer-linguistic components can therefore only be answered empirically. This statement is the specific form of a general information science rule: during a machine-supported information process, single components do not achieve what their function description expresses, but what they bring about in an overall context.

Every change in the text basis therefore calls for new empirical tests. The same applies to a change in other parameters such as user groups, application object, etc.

It is impossible to maintain some widespread spurious plausibilities which assume that an improved (linguistic) performance could at most have no effect, whereby poorer results for richer content analysis methods in terms of computer linguistics would be excluded from the outset. This attitude will at least remain incorrect as long as a complete mechanical simulation of interpersonal cognitive information behavior is not available, for whatever reason. The use of subcomponents, however, leads to the field of restricted systems with their own laws.

4. Statistically oriented methods in information retrieval

Whereas computer-linguistic methods use traditional free-text systems when reducing content analysis, but leave the standard model of IR largely untouched, quantitative-statistical methods (under designations such as best-match or nearest-neighbour methods with or without relevance feedback) change the retrieval process in a more far-reaching manner. They are primarily regarded as techniques against the following negative properties of Boolean retrieval in which user-oriented and cognitive aspects (cf. section 2) are for the most part ignored.

• Boolean retrieval divides the document set - without any interim stage - into two discrete subsets: into documents which fulfill the "exact match" (= relevant documents) and those which do not. Documents with three found terms are rejected in a search query comprising four terms linked by the AND operator in exactly the same way as those with 0.

All outputted documents are equivalent from a system viewpoint. The last document on the results list can best satisfy the information requirements of the user. At the descriptor level, this corresponds to the impulse to use all descriptors as "equally important", something which users regard as an inadmissible simplification.

• Users often have problems in making adequate use of the logical operators AND, OR and NOT. One reason for this is that the semantics of the logical operators do not agree with the semantics of the natural language terms. The priorities of the Boolean operators must also be known.

Commercial developments in statistical systems are only slowly starting to compete with traditional methods on the market. One example is TOPIC (cf. Wood & Moore, 1993 for an overview of commercially available IR software). There are also several experimental system variants which are already being used at universities or scientific institutes for real applications (e.g. the INQUERY system of the University of Massachusetts, Broglio & Croft, 1994).

4.1 Basic common features: ranking of the results list and descriptor sequence

With regard to their theoretical background, non-Boolean retrieval models can be divided into probabilistic (statistical probability theory), vectorial (vector space model) and fuzzy retrieval models (theory of inexact quantities) which interpret the similarity function in different ways. As the TREC studies (Harman, 1993) showed, the theoretical differences have practically no effect on the retrieval results, which is why, in my opinion, it is possible to make use of the basic architecture common to all approaches in application development.

Best-match methods can be characterized by the fact that the user strings descriptors together during the query without using Boolean operators and the most relevant documents should come at the start of the results list. This so-called ranking is generated by the system based on similarity criteria.

4.2 Determination of similarity between query and document

The similarities determined by the system define the ranking of the document on the results list. The most widespread similarity is the so-called "vector dot product" in which the similarity is calculated from the product sum of the (weighting) of the terms which appear in the query and the document. The higher the calculated value, the higher the document on the results list.

Also used are the Cosine measure (normalized, including the document length), the Dice coefficient and the Tanimoto measure (cf. Ruge, 1994a). A minimum similarity is often determined by a certain threshold value or the number of required documents defined by the user is utilized as a limitation criterion.

4.3 Weighting

Weighting of the terms in the documents is normally included in the similarity scale. Weighting is automatically determined for every term in a document in relation to certain quantitative properties of the document or the collection of documents. For example, the number of documents in database n and the frequency of the term t in the document collection are included in the "inverse document frequency" weighting. The equation G=log(N/F(t)) means that general terms (quantitative characteristic: high frequency) contribute less to the relevance of a term than specific terms which seldom occur. The yardstick can also include the frequency of a term in a document (within term frequency). The more often t occurs in a document, the higher its weighting, and the less often it appears in other documents too, the better.

A possible connection with the computer-linguistic approaches is already apparent here. All methods mentioned in section 3 can be defined as pre-determination of t. This ensures, for example, that singular and plural forms are not included as two different terms in the calculations.

Some weighting measures also take account of the number of the different terms within a document and/or specify limits for the occurrence frequency of a term (e.g. t must occur at least three times in the data collection). It is also obvious to introduce weighting rules relating to text types, e.g. to weight terms in headings higher than others. In these variants which have to be determined according to their individual application and whose impact on ranking must be verified empirically, there appears to be great potential for improvement with approaches based on non-Boolean retrieval. Generally speaking, however, this applies both to all quantitative-statistical methods and computer-linguistic algorithms. We must deal with restricted system whose real efficiency is exclusively

empirical and can only be proved in relation to specific user groups and application situations.

Wider use of automatic weighting methods in commercial mass databases is primarily prevented by the fact that the term weightings must be recalculated with every change in the data set.

Weightings cannot only be linked to the document terms. The query term can (also) be weighted. As a rule, the user himself (herself) determines intellectually the weighting which he (she) wants to give his query term in his (her) descriptor list (see, however, relevance feedback).

4.4 Relevance feedback

This method calls for a query with at least two stages and the cooperation of the user. He (she) evaluates the results list by crossing the item in, for example INQUERY if an output document was of "relevance" to him (her). The system therefore knows that it is "correct" and uses this dynamic control knowledge from the current dialog situation to "recalculate" the original query. Query terms, which occur frequently in documents specified as "relevant", are given a higher weighting during this reformulation of the query or they are added to the original query. As a result, the modified results list will better satisfy the information requirements of the user (cf. Robertson & Sparck Jones, 1976).

4.5 Clustering and extension lists

The objective of cluster methods is to divide document sets or their descriptor lists into classes whose members are closely related in terms of contents. The degree of relationship is measured formally by means of the joint occurrence of the descriptors in the document. Various mathematical methods are used in this case (cf. Salton & McGill, 1987). A document cluster therefore contains documents whose related descriptors agree as far as possible with one another. A centroid is determined for every cluster and is regarded almost as the "most typical" representative of the cluster. Descriptor clustering is based on the same basic idea: a descriptor t1 is regarded as closely associated with t2 in terms of contents if it occurs as often as possible together with t1 in the documents in the database. In both cases, the cluster is used to determine relevance. If a document d or a descriptor t is regarded as relevant to the information requirements of the user, the members of its cluster are also regarded as relevant (cf. Sparck Jones & Van Rijsbergen, 1973; Croft, 1980).

Cluster techniques supplement the methods already discussed. Document clustering is hardly used on account of the large amount of time spent on mathematical calculations, especially for updating application-related systems. However, descriptor clustering is used. This includes, for example, the extension list of REALIST (Ruge; 1992; Schwarz & Thurmair 1986). Extension lists are networks of terms which correlate statistically with the starting term. Figure 1 shows a simple example of the term *CPU* and how often (in %) the term *CPU* occurs with other terms in documents (number).

Extension lists give rise to the following working method:

• The descriptors are extracted in an initial step (by means of any IR system such as STAIRS or GOLEM/PASSAT).

• Extension tables are then drawn up based on terms and their occurrence, followed by calculation of the correlations between the descriptors of a document and those of other documents in the data set.

The extension lists may be made available to the user as an additional tool for independent strategy planning (research assistance) or may also act as a basis for automatic research expansion as part of IR components (see, for example, Grefenstette, 1992).

The preparation of extension lists normally does not depend on language. In the REALIST context, extension lists were prepared based on a representative subquantity of the total document set and then transferred to the overall stock.

CPU		in document: 105 correlation terms: 1817	
correlation (%)	Term	correlation (%)	Term
65.71	DATA	7.62	LOGIC
49.52	MEMORY	7.62	LOCATION
33.33	STORE	7.62	COMMUNICA- TION
32.38	CENTRAL	7.62	CHIP
...			

Figure 1: Use of term CPU

4.6 Extended Boolean retrieval

As with computer-linguistic and quantitative-statistical methods, the best-match method and Boolean retrieval also offer a mixed form. The main starting point for this consideration were the empirical observations that both search methods lead to widely differing result quantities and that users of non-Boolean systems regarded the lack of Boolean operators in certain situations as a handicap. They want, for example, to link synonyms by OR. With extended Boolean retrieval, the user starts with a conventional Boolean query which will, however, only determine a preliminary selection of potentially relevant documents from the search strategy. A ranking method is then added to the query in order to arrange the preliminary selection according to actual relevance (cf. Salton et al; 1983, Bookstein, 1981).

In order to ascertain the correctness of the basic conviction that the user's objective - "preliminary selection" - substantially reduces the disadvantages mentioned at the start of section 4, empirical calculations would have to be made for specific application areas. The advantage would have to be so great that it exceeds the higher cognitive burden on the user, which is automatically produced due to concept doubling.

5. Development potential

According to the considerations in sections 1 to 4, content analysis produces three main groups of potential starting points for improving the retrieval performance of current commercial systems based on Boolean algebra which promise to be successful in a relatively short space of time and with limited expenditure.

5.1 Additional computer-linguistic modules

Traditional free-text retrieval reduces content analysis to a symbol-oriented analysis. Documents are regarded as the stringing together of chains of symbols which are separated by blank spaces or punctuation marks. The erroneous matching processes thus produced can be reduced by using computer-linguistic methods. Even if empirical tests are ultimately the only way to find out whether the retrieval performance of a specific application area can be improved substantially by computer-linguistic methods, there are many plus points as regards improved quality.

Computer-linguistic components are now available as fully developed basic software both in the scientific and commercial sectors (examples SOFTEX and GERTWOL).

5.2 Use of quantitative-statistical methods

Various individual components appear highly promising as regards short-term development of current commercial systems based on Boolean algebra.

a) REALIST (Retrieval Aids by Linguistics and Statistics) contains statistical techniques (extension lists) which lead to terms which correlate with a descriptor sought in the database. The user can select or exclude additional terms from the extension list (to prevent an unwanted reference of the selected descriptor) and thus specify his (her) information requirements precisely. This method is well-suited conceptually to linguists and also counters the restriction on Boolean queries, namely that the terms cannot be weighted.

b) Due to the anticipated performance problems with quantitative-statistical methods, tests using an extended Boolean retrieval model appear promising. Thanks to this model, the Boolean query can be carried out using a traditional system such as STAIRS. The thesis is that Boolean logic would no longer be normally deployed by the user for complicated links, but merely to determine a pre-selection. A ranking algorithm, which works with weighted document terms, could then be added to the resulting results quantity. The empirical calculation of specific parameters from the specific application context appears to be a highly promising aspect. As additional "references", these parameters reveal the significance of descriptors (e.g. highlighted position of a term in the title or basic principle of a document).

Based on the information in b), the entire Boolean query could be replaced experimentally in a second development phase by a ranking method with relevance feedback methods. Whether this will lead to improved quality in a specific application and what calculation methods should be selected can only be determined by means of empirical tests using suitable prototypes.

5.3 Summary

The methods discussed in this paper therefore provide a sufficient number of detailed approaches which could improve the information performances of current commercial databases based on Boolean algebra without forcing a complete new start. It would be possible to combine them by integrating additional research strategies, integrating application-oriented components of an intelligent IR system and rearranging the user interface of such a system. The latter aspects had to be excluded from this article due to reasons of

space together with the increasingly urgent problems of connecting text and factual IR systems.

Literature

Bates, M.J. (1989): The design of browsing and berrypicking techniques for the online search interface. *Online Review 13*:407-424.

Biebricher, P., Fuhr, N. & Niewelt, B. (1986): Der AIR-Retrievaltest. In: Lustig, G. (Hrsg.): *Automatische Indexierung zwischen Forschung und Anwendung* (pp. 127-143). Hildesheim: Olms.

Belkin, N.J. (1993): Interaction with Texts: Information Retrieval as Information Seeking Behavior. In: Knorz, F., Krause J. & Womser-Hacker, C. (eds.): *Information Retrieval '93*. Proc. 1st GI-Conference on Information Retrieval. Konstanz: Universitätsverlag.

Bookstein, A. (1981): A comparison of two systems of weighted Boolean retrieval. *Journal of the American Society for Information Science 32*:275-279.

Broglio, J., Callan, J. & Croft, W.B. (1994): INQUERY System Overview. In: *TIPSTER text program phase 1: Proceedings of a Workshop held at Fredericksburg, Virginia* (pp. 47-67). San Francisco: Morgan Kaufmann.

Croft, W.B. (1980): A Model of Cluster Searching Based on Classification. *Information Systems*, Vol. 5, No. 3:189-195.

DATEV 1994: Datenbanken Anwenderhandbuch. Nürnberg: DATEV.

Fuhr, N. (1988): *Probabilistisches Indexing und Retrieval*. Dissertation, TH Darmstadt, Fachbereich Informatik.

Grefenstette, G. (1992): Use of syntactic context to produce term association lists for text retrieval. In: Belkin, Nicholas et al.: *Proceedings of the 15th Annual ACM SIGIR* Conference on Research and Development in Information Retrieval, SIGIR '92 Copenhagen (pp. 89-97). New York: ACM Press.

Haenelt, K. (1994): Das Textanalysesystem KONTEXT. Konzeption und Anwendungsmöglichkeiten. *Sprache und Datenverarbeitung*. Bd. 18, No. 1:17-31.

Harman, D. (Ed.) (1993): *The First Text Retrieval Conference (TREC-1).* Gaithersburg, National Institute of Standards and Technology. Special Publication (pp. 200-207). Springfield: NIST.

Hitzenberger, L. (1987): Phonological Access to Databases. In: Luelsdorff, P. (ed.): *Orthography and Phonology* (pp. 65-76). Amsterdam: Benjamins.

Informationszentrum Sozialwissenschaften (IZ): Jahresbericht 1994. Bonn: IZ.

Knorz, G. (1994): Automatische Indexierung. In: Hennings, R.-D. et al. (Hrsg.): *Wissensrepräsentation und Information-Retrieval* (pp. 138-198). Modellversuch BETID Lehrmaterialien; 3. Potsdam: Universitätsverlag.

Krause, J. (Hrsg.) (1987): *Inhaltserschließung von Massendaten*. Hildesheim: Olms.

Krause, J. (1990): *Zur Architektur von WING: Modellaufbau, Grundtypen der Informationssuche und Integration der Komponenten eines Intelligenten Information Retrieval*. WING-IIR-Arbeitsbericht. Regensburg: Informationswissenschaft.

Krause, J. (1992): Intelligentes Information Retrieval: Rückblick, Bestandsaufnahme und Realisierungschancen. In: Kuhlen, R. (Hrsg*.): Experimentelles und praktisches Information Retrieval* (pp. 35-58). Festschrift für Gerhard Lustig. Konstanz: Universitätsverlag.

Krause, J. & Womser-Hacker, C. (Hrsg.) (1990): *Das Deutsche Patentinformationssystem*. Köln: Heymann.

Kuhlen, R. (1985): Verarbeitung von Daten, Repräsentation von Wissen, Erarbeitung von Information. Primat der Pragmatik bei informationeller Sprachverarbeitung. In: Endres-Niggemeyer, B. & Krause, J. (eds.): *Sprachverarbeitung in Information und Dokumentation* (pp.1-22). Berlin: Springer.

Kuhlen, R. (1991*): Hypertext. Ein nicht-lineares Medium zwischen Buch und Wissensbank*. Berlin: Springer.

LINGSOFT (1994): GERTWOL. Questionnaire for MORPHOLYMPICS 1994. LDV-Forum 11 (1). *Sonderheft MORPHOLYMPICS*: 17-29.

Lustig, G. (1986): Eine anwendungsorientierte Konzeption der automatischen Indexierung. In: Lustig, G. (Hrsg*.): Automatische Indexierung zwischen Forschung und Anwendung* (pp. 1-12). Hildesheim: Olms.

Möller, Tong (1993): *Juris für Juristen*. (Law for Jurists). Dissertation, Universität des Saarlandes.

Rahmstorf, G. (1994): Semantisches Information Retrieval. In: Neubauer, W. (Hrsg.): *Proceedings Deutscher Dokumentartag* 1994 (pp. 237-260). Trier: Universität.

Robertson, S. E. & Sparck Jones, K. (1976): Relevance Weighting of Search Terms. *Journal of the American Society of Information Science*, Vol. 27:129-146.

Ruge, G. (1992): Experiments on linguistically-based term associations*. Information Processing & Management*, Volume 28, No. 3: 317-332.

Ruge, G. (1994a): *Wortbedeutung und Termassoziation. Methoden zur automatischen semantischen Klassifikation*. Dissertation, TU München, Fakultät für Informatik.

Ruge, G. (1994b): *Skript Tutorial Computerlinguistik*. 1st GI-Conference on Information Retrieval, Regensburg, Sept. 1993.

Salton, G. (Ed.) (1971): *The SMART Retrieval System*. Experiments in Automatic Document Processing. Englewood Cliffs: Prentice Hall.

Salton, G., Fox, E. & Wu, H. (1983): Extended Boolean Information Retrieval. *Communications of the Association for Computing Machinery* 26:1022-1036.

Salton, G. & McGill, M.J. (1987): *Information Retrieval*. *Grundlegendes für Informationswissenschaftler*. Hamburg: McGraw-Hill.

Schneider, C. (1987): Analyse der Texterschließung. In: Krause, J. (Hrsg.): *Inhaltserschließung von Massendaten* (pp. 56-65). Hildesheim: Olms.

Schwarz, Ch. & Thurmair, G. (Hrsg.) (1986): Informationslinguistische Texterschließung. Hildesheim: Olms.

Sparck Jones, K. & Van Rijsbergen, C.J. (1973): A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, Vol. 29:251-257.

Weizenbaum, J. (1976): *Computer Power and Human Reason*. From judgement to calculation. San Francisco: Freemann.

Wood, Joanna & Moore, Caroline (1993): *European Directory of Text Retrieval Software.* Aldershot: Gower.

Wolf, Gerhard (1992*):* JURIS - Ein denkbarer einfacher Zugang zu allen Informationen, die Sie brauchen*. jur-pc. 4*: 1524-1810.

Womser-Hacker, C. (1989): *Der PADOK-Retrievaltest. Zur Methode und Verwendung statistischer Verfahren bei der Bewertung von Information-Retrieval-Systemen*. Hildesheim: Olms.

Zimmermann, Harald (1993): *Grundlagen und Verwendung der linguistischen Software von Softex.* Saarbrücken: Softex GmbH.

Address:

Professor Dr. Jürgen Krause, Informationszentrum Sozialwissenschaften, Lennéstr. 30, D-53113 Bonn, Germany, Tel.: +49-228/228-1145, Fax: +49-228/228-1120, e-mail: Krause@IZ-Bonn.GESIS.d400.de