

relevancy

White  
Paper



# *Making information useable*

*by Peter Schäuble and Martin Braschler  
May 2002*

Tel. +41 43 255 25 25  
Fax +41 43 255 25 26  
info@eurospider.com  
www.eurospider.com

Eurospider Information Technology AG  
Schaffhauserstrasse 18  
CH-8006 Zürich  
Switzerland

## *Inhaltsverzeichnis*

1	Introduction.....	3
2	Indexing of information objects .....	5
2.1	Simple indexing methods .....	6
2.2	Advanced indexing methods.....	6
2.3	Innovative indexing methods.....	8
3	Matching of information objects.....	9
3.1	Simple matching methods for information objects.....	9
3.2	Advanced matching methods for information objects .....	9
3.3	Innovative matching methods for information objects .....	10
4	Outlook.....	11

## 1 Introduction

The core competence of Eurospider Information Technology AG lies in the area of the so-called Information Retrieval. This technical area came into existence more than 40 years ago with the goal of "*Making documents automatically accessible according to content criteria*". The discipline of Information Retrieval, which was largely academic at the beginning, has gained considerably in commercial importance with the transition into the information age.

The availability of business critical information has increased significantly for a number of reasons (World Wide Web, Mail, CRM and ERP systems). In addition, there is also an even more rapidly growing flood of data that makes the finding of relevant information more difficult. At the same time, new and varied application possibilities for Information Retrieval techniques have arisen, which have created the need for corresponding systems.

Traditionally, an Information Retrieval system consists of two main components: Indexing of the content (Indexing) and Matching. By *Indexing of Information Objects*, the object is to bring these into a form that makes an effective comparison possible. This is required because information objects are a priori not directly comparable, for example, due to the many possibilities of describing the same facts in natural speech. Indexing makes possible:

- the integration of different information objects, such as documents, search queries, user or interest profiles, subjects of a taxonomy, and others
- the integration of different, heterogeneous sources
- the extraction of a maximum of information from these objects for use in later comparisons

By *Matching of Information Objects*, it needs to be determined whether a semantic exists between the objects. The result of comparisons can be a Yes/No decision, but could also be a numerical value that expresses the strength of the semantic relationship. The matching of information objects can be used for many different purposes:

- Matching between a search query and a document should indicate whether the document is relevant to the query, and, consequently, should be presented to the user as a component of the search result. The result of this comparison is a value that expresses how relevant the document is to the query, and thereby determines how near to the top of the search result the document will be placed.
- Matching of a user/interest profile with a document should indicate whether the document should be forwarded to the user.

- Matching of a subject of a taxonomy (theme catalogue) with a document should indicate whether the subject should be assigned to the document. In the case of a catalogue, the document will then be listed in the corresponding category.

Different methods have been developed for indexing and matching of information objects. Information Retrieval systems can be differentiated by the simple, advanced or innovative methods used for the indexing and matching.

The effectiveness of these methods can be objectively measured by variables such as *Recall* and *Precision*. The Recall expresses the proportion of the desired information that will be found, while the Precision measures how much of the found information is relevant.

There now follow several examples of the combination of specific methods for the indexing and the comparison.

#### **Example 1: News Channels**

Typical news portal with educable channels based on a *simple* indexing and using neither language analysis nor document structures. In this case, for example, „share issues“ and „New issues of shares“ are not comparable, and identical numbers lead to undesired hits, for example, when a telephone number is compared with a car registration number. The educable channels of news portals of this kind are based on *advanced* statistical methods for the comparison (e.g., Bayes' Classifier or kNN), which decide whether a new document should be fed into a channel on the basis of training examples.

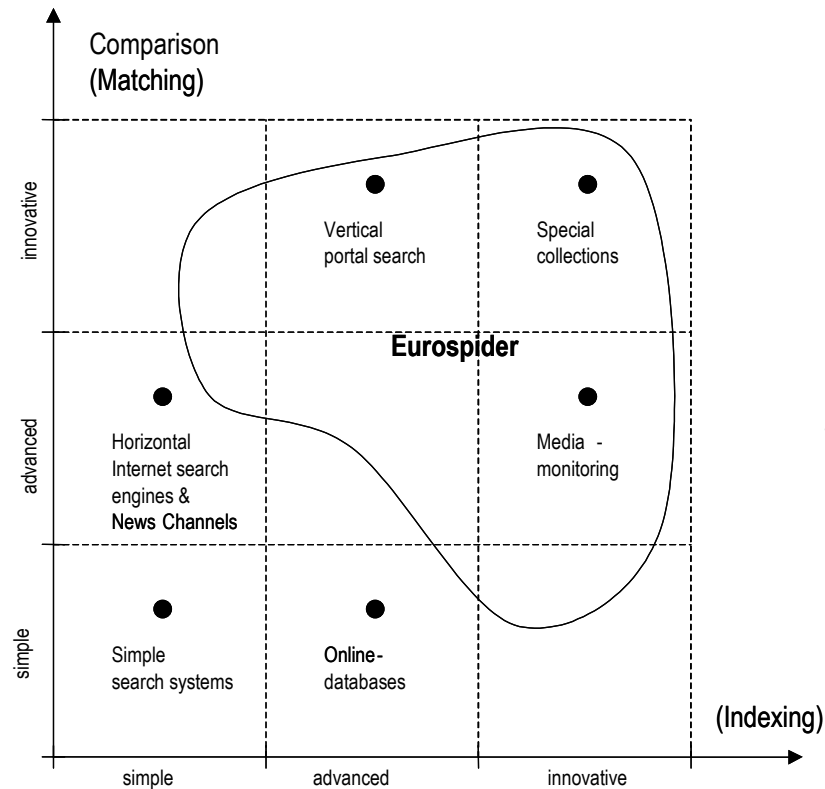
#### **Example 2: (Horizontal) Internet Search Engine**

Internet search engines index a very large number of extremely heterogeneous Web pages, so that even the best search engines use *simple* indexing methods. To do this, *advanced* matching methods come into use, which, for example, take the hyperlink structure of the Web into account (Page Rank, HITS algorithm, SAE, etc.).

#### **Example 3: Special collections of documents**

Special subject-specific and homogeneous document collections require *complex* indexing with automatic recognition of references/Meta-data and the inclusion of special technical vocabulary. At the same time, a *complex* matching method is also required, for example, to guarantee a multi-lingual access to the relevant documents that are formulated in a language different to that used for the search query.

The graph below illustrates the position of the examples mentioned and of other examples, as well as the positioning of Eurospider, which has concentrated on advanced and innovative methods, such as vertical portal searching, special collections and media monitoring.



In the following paragraphs, the two traditional retrieval components – indexing and matching – will be discussed in detail (Paragraphs 2 and 3). The Outlook (Paragraph 4) contains information on how Information Retrieval systems could look in the future.

## 2 Indexing of information objects

For humans, the decision of whether a document is relevant to a search query is a very complex process, which is subject to numerous uncertainties. The same high level of complexity is found in the decision of whether a document matches a user profile or belongs to a certain subject. In an Information Retrieval System, this decision is prepared in that the information objects are indexed, i.e., are made comparable. The indexing comprises:

- Reduction of information, i.e., unimportant parts are left out, such as stop words (the, in, a, etc.)
- Normalization of information, for example, by language analysis (orders = order) or normalization of numbers (11 Mio. Fr., CHF 0.011 Bio., eleven million Francs = CHF 11 000 000.–).
- Enrichment of the information: assign subjects or detect entities (persons, companies, products, etc.).

The following indexing methods have, among other things, proved useful in making information objects comparable.

## 2.1 Simple indexing methods

### (1) Conversion of document formats

Documents in all standard Office formats (Word, Excel, PowerPoint, Lotus Notes, WordPerfect etc.) and the important exchange formats (HTML, XML, SGML, Postscript and PDF) will be suitably converted so that they can be read into the system.

### (2) Conversion of fonts

Various coding systems for textual information (ASCII, ANSI/ Windows, ISO Latin, KOI8 Cyrillic, etc.) will be converted into a suitable internal format so that they can be processed by the system.

### (3) Word-segmentation

The individual words will be extracted from the information object. In doing this, sentence punctuation and spaces, among other things, will be removed.

### (4) Stop-word elimination

Certain very common words (articles, prepositions etc.) will be eliminated. These words do not help in distinguishing relevant from non-relevant information. Through the elimination, the size of the index will also be reduced and comparisons will be accelerated.

### (5) Parsing information objects

An information object will be searched through in order to identify those parts that should be indexed. Other text parts (certain formatting codes, etc.) will be ignored.

## 2.2 Advanced indexing methods

### (6) Stemming

In natural spoken language, expressions are used in various word forms, depending on their use in the grammatical constructions. In order to ensure that as much relevant information as possible will be found, words have to be normalised so that words that are not given in exactly the same form will still be given as hits during matching with search expressions.

(7) Decomponding

Some languages, such as German, permit the formation of complex expressions by joining a number of simple words together without any intermediate spaces. Compound words of this kind can often be written in a different manner, however, or are frequently only partially referenced in a query. It is therefore important to break them down into their constituent parts.

(8) N-Gram

Indexing by word is suitable when a system should process documents with few or no typing or grammatical errors and if the documents have been registered in a language that is known to the system (necessary for stemming/decompounding). If this is not the case, the system can break down alternative words into smaller units ("N-Gram"), which makes an error-tolerant matching possible.

(9) Fuzzy Matching

Fuzzy Matching enables robust retrieval of relevant information, especially in the case of typing errors and alternative spellings. Fuzzy Matching generates relevant matches independent of whether the query or the document contains the misspellings or alternative transcriptions and transliterations.

(10) Language detection

Stemming and decompounding are usually dependent on the language. If a system has to process documents in different languages, it is necessary that it must first detect the language for each document, and possibly also for each paragraph, or even for each sentence.

(11) Statistical text categorisation

It is increasingly becoming the task of an Information Retrieval System to not only search through large quantities of search expressions, but to automatically fit documents into a hierarchy of categories that have been defined through complex criteria. Statistical procedures solve this problem on the basis of training examples.

(12) Structured information objects

Structure in information objects is detected and evaluated in order to later allow specific access to information in certain fields only.

## 2.3 Innovative indexing methods

### (13) Concept sensors

Concept sensors permit the formulation of very complex relationships that require the integration of extensive rules. With their help, it is possible to detect facts that only arise through certain correct combinations of several factors.

### (14) Entity recognition

In many cases, background knowledge is necessary in order to make optimal use of information. Entity recognition identifies words as names (of persons, companies, locations) and thereby makes it possible to place these in connection with additional information.

### (15) Nominal phrase extraction

A combination of several words can often have a more specific meaning than the sum of its individual components. Phrases, i.e., expressions involving several words, are recognised and are processed as a unit.

In connection with the indexing of information objects, Eurospider has the following USPs (Unique Selling Propositions).

The relevancy system contains an *indexing pipeline* that permits simple to innovative indexing methods to be arranged one after another in steps, which can make use of the results of the previous steps. Thanks to the flexibility of the pipeline, powerful customised indexing can be realised very simply.

In order to categorise documents, the Information Retrieval System must decide whether the document matches a user/interest profile and/or whether the document belongs to a subject in a taxonomy or not. The system must make a decision, as, where appropriate, the document will be transmitted to the user and/or the document will have to be listed under a subject. As even the best statistical categorisation methods only have a limited accuracy, Eurospider developed so-called *Concept Sensors* that are based on a rule-based approach and make high precision categorisation possible. The relevancy system has both statistical categorisation and concept sensors, so that the optimal method can be used for each application.

### 3 Matching of information objects

After indexing, the information objects are compared, in order to match search queries, user/interest profiles and subject of a taxonomy with relevant information. Matches of this kind are associated with numerous uncertainties. Whether a document satisfies a user interest depends, among other things, on the specialised knowledge that is assumed. A specialist may make a different decision than a layman. An Information Retrieval System is therefore confronted with an "insoluble" problem, which has to be solved as well as possible in as many cases as possible. The different methods for the matching of information objects will be briefly described in the following.

#### 3.1 Simple matching methods for information objects

##### (16) Boolean Retrieval

The so-called Boolean Operators (AND, OR, NOT) are used to explicitly define relationships between individual search terms.

##### (17) Coordination Level Matching

The ranked list is sub-divided into individual sections, which are arranged according to the number of search terms found.

#### 3.2 Advanced matching methods for information objects

##### (18) Probabilistic

Ranked lists are sorted on the basis of estimates of the probability that an object is relevant. There are sophisticated formulae for the calculation of the probabilities.

##### (19) Rule-based

Information objects are organised with the help of a number of rules, so that they can then be presented in a ranked list. This procedure permits a simple adaptation to customer-specific ranking wishes.

##### (20) Relevance feedback

The user can evaluate search results based on their relevance, whereby the system automatically refines the query further and supplies better search results.

##### (21) Query expansion

It is often difficult for the user to determine the terminology with which a sought fact will be handled in the available documents. Automatic query expansion widens search queries by the terms used, and thereby helps to achieve more comprehensive search results.

(22) Metadata

Even when they are only partly structured, many information objects contain metadata, which considerably facilitate the access. This kind of metadata (date, source, etc.) can be used to exclude irrelevant information.

(23) Sub-collections

Documents can often be divided into sub-collections on the basis of source, thematic or other criteria. As a result, a user can then specifically enable and disable individual parts of the document collection.

(24) Duplicate elimination

Certain document collections contain a large amount of redundancy and, in particular, many objects that appear many times in exactly the same, or almost identical forms. These duplicates must be grouped together and be displayed to the user in a more compact form.

(25) Access limitation

Especially in companies, information is often only approved for distribution in a restricted manner. A search system must avoid providing unauthorised users with knowledge of or even access to secret information through a “back door”.

### 3.3 Innovative matching methods for information objects

(26) Cross-language retrieval

In today's world of globalisation and multi-national organisations and companies, it is becoming increasingly common to index document collections that contain objects in many different languages. It must be possible to efficiently access such collections using only a single query formulated in the language preferred by the user.

(27) Retrieval of passages

In longer information objects, it is often the case that only short sections are relevant for the answering of a search query. In order to deliver a good search result, the system must be able to identify and correspondingly weigh such sections.

In connection with the matching of information objects, Eurospider has the following USPs available: The relevancy system contains a *large spectrum of matching methods*, and can be adapted individually to customer needs. In *multi-lingual search*, Eurospider benefits from their many years of research activity and regular participation in international competitions.

#### 4 Outlook

In the pioneering area of Information Retrieval (1960-1980), the task was basically to find as much relevant data, and only relevant data, as possible within a well-defined document collection in order to write a report. The situation has now changed fundamentally: today, a very wide spectrum of tasks is considered, such as

- Initiating transactions (booking cheaper holidays, buying software, etc.)
- Decision support (preparing market information, Compliance Management, etc.)
- Retrieval of facts (contact information, specifications, etc.)
- Interactions (E-Government, Online applications, etc.)
- Co-operation (projects, virtual companies, etc.)

Today, it is therefore not only a matter of preparing useful information, but also of preparing these in a *personalised* and *task-specific manner*.

The Information Age places new and higher demands on Information Retrieval Systems, which are no longer satisfied by the two traditional components of indexing and matching alone. In specialist circles, new components are being intensively discussed, such as machine translation, information visualisation or image and language recognition. With regard to Eurospider's goal of making information useable, the benefits of these new components cannot be measured using the traditional performance indicators of Recall and Precision.

The probability ranking principle of Cooper and Robertson justifies the so-called Relevance Ranking, in which the documents are sorted by relevance, and are then presented to the user. This principle will lose its importance as the optimal approach for its original task. A Relevance Ranking can no longer be regarded as the optimal procedure in every case. In the same way, the measurements of the Retrieval Effectiveness - expressed in Recall and Precision - that are associated with it will also become less important. New, task-specific measures of utility are being increasingly used, and a completely new Information Retrieval paradigm is required that can optimise these new measures of utility.

Eurospider was one of the first companies to place the future requirements and the new components in a generalised framework<sup>1</sup>. In doing this, it must be assumed that there is a single person, who has a task to solve. Whether the person can solve this task both quickly and well in terms of quality and quantity in the Information Age depends on whether useful information is

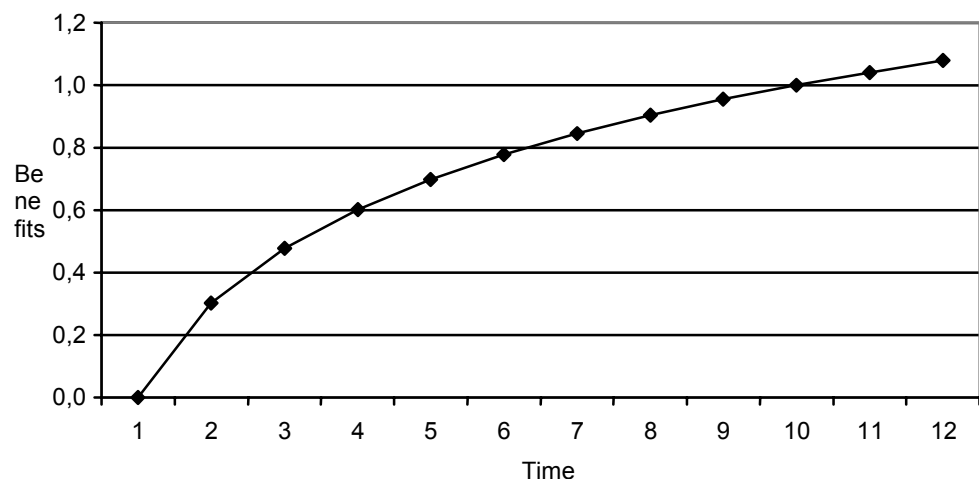
---

<sup>1</sup> Digital Library Conference, Berkeley 1999.

available to him. The benefits of the information in this case are both task and person-specific, and depend on various factors:

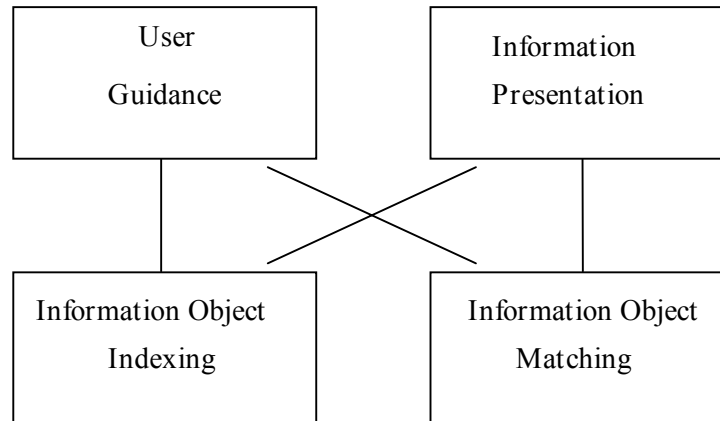
- How much time is needed to read (text), listen to (audio) or view (images, video) the information?
- How much time is needed to understand the information?
- How high are the costs of the information?
- How good is the quality of the information?
- How up-to-date is the information?
- How reliable is the information?

Even if the benefits of these functions are only partly understood nowadays, Eurospider is working on the assumption that future Information Retrieval Systems will not only have to optimise Recall and Precision, but such benefits also. In concrete terms, this means that the greatest possible benefit should be realised in as short a time as possible, specific to both the task and the person.



In this new, generalised framework, it is once again two components that play a decisive role: user guidance and information presentation. User guidance assists the user in the interaction with the Retrieval System by selecting an action in the next step that will bring the greatest possible benefit. After carrying out the action (e.g., modifying a search, translating and grouping together the search result, relevance feedback, etc.), the information must be presented in such a way that, once again, the greatest possible benefit results. The interplay between user guidance and information presentation should result in a benefit function that, depending on time, should grow as quickly as possible (see graph above).

These two new components correspond to the well-known components Indexing and Matching. In this way, the path towards a new generation of Information Retrieval Systems has been prepared.



User Guidance includes virtual retrieval experts that provide the user with specific recommendations. Recommendation systems of this kind can be based on methods such as Cooperative Filtering or Data Mining. The presentation of information also opens up some very interesting perspectives. Methods for grouping documents together, automatically translating them and extracting answers from them will be able to be used for these purposes in the near future.

The Vision of Eurospider is to make information useable. Eurospider will therefore continue to analyse, develop and make specific use of promising technologies in the future.

Eurospider is the leading Swiss expert for relevancy retrieval – the retrieval and preparation of relevant information using state-of-the-art methods. With its relevancy software, Eurospider realizes business-critical solutions in the field of information retrieval.