Information Need Assessment in Information Retrieval
Beyond Lists and Queries

Frank Wissbrock

Department of Computer Science
Paderborn University, Germany
frankw@upb.de

Abstract. The goal of every information retrieval (IR) system is to deliver relevant documents to an user's information need (IN). Therefore an accurate IN assessment is essential to the quality of the system's search results. However, many IR systems ask the users to assess their information needs and communicate them to the system, usually in form of queries. The systems assume the queries to be a perfect assessment of the information needs and deliver relevant information, ending the interaction. However, experiences showed that in many cases the information need cannot be specified in a single query.
This paper addresses the problems of simple IN assessment and proposes a multi-interface IR system to overcome the problems. Such a system supports the user with several search interfaces for different search contexts. Exemplarily the document retrieval engine AiSearch from the Knowledge-based Systems Group at Paderborn University is reviewed to demonstrate some interfaces. This includes a cluster-based interface, a concept taxonomy interface, and a chronological document relations interface.

1 Introduction

Information need (IN) is one of the most important concepts in information retrieval (IR) theory. It is the main input parameter for most IR operations as well as the main evaluation criteria for the quality of the delivered information. But even though the concept of information need is central to the success of any IR system, most IR models treat the concept as intuitively clear and informal. From this viewpoint the importance of information need assessment is often underestimated. Indeed in most IR systems information need assessment is user business. Take for example common internet search engines. They require the users to formulate their information needs in form of a query, assuming that the query is an accurate definition of the information need. However, it was shown that this assumption does not hold for many IR transactions [1] [2].

Starting from the viewpoint that common search engine interfaces do not support an accurate information need assessment this paper proposes an IR sytem with multiple user interfaces, where each of the interfaces fits a certain search context of the user. Based on a theoretical and historical discussion of IN assessment in section 2-4 the multi-interface model is presented in section 4. Section 5 describes AiSearch, a search engine project of the Knowledge-based Systems Group at Paderborn University, to demonstrate how parts of the model were implemented and how they look like. [3].

2 Historical Developments in Information Need Assessment

Before a formal definition of information need and informantion need assessment is given some approaches to information need assessment are briefly reviewed in their historical context. The intention is to build a foundation for the definitions given in the next section.

2.1 Query approach

The query approach was the first IN assessment method and is still widely used. It was developed in the late 1950s and early 1960s in the context of text properties research and the formulation of the standard IR model [4] [5]. The basic idea of the approach is to let the user assess his information need. Therefore the user enters a query, which usually consists of one or more natural language terms. In turn the system presents all documents from its database that match the query. In 1965 Roccio added an additional step to the query approach: the relevance feedback [6]. With relevance feedback the user judges the result in light of its relevance to his or her information need. Therefore he classifies the returned documents into two classes, the relevant documents and the non-relevant documents. After that the system uses the classification to adjust the initial query and the retrieval process starts again with the adjusted query. The new result is, if necessary, classified again by the user. The assessment is repeated until the query is a perfect representation of the user's information need.

2.2 Dialog approach

The query approach bases on the assumption that the user knows what his information need is and that he can adequately communicate it to the system. Relevance feedback takes care of an accurate IN assessment. However, relevance feedback implicitly assumes that the information need itself stays constant over time, even when the user has gained new knowledge during the search process. Recognizing that this assumptions did not hold always, Oddy proposed a dialog interface in 1977 [1]. The basic idea is that a user's understanding of his information need underlies a continuing evolution while new information is retrieved. The dialog interface allows the user to reformulate his previous query to broaden or narrow the retrieved information or to shift the search goal. The interaction is continued until the needed information is found. The difference to the query approach is that Oddy embedds the user into the IR system. The user is no longer only an input giver but a part of the retrieval process.

Some years later Belkin shifted the focus even farther to the user and his information need [2]. He asked why most users are not able to specify their information needs in an appropriate way. The answer was given by a new element in the user model: the "anomalous state of knowledge" (ASK) of the user [2]. Therefore every user who faces a problem or situation has a feeling about a gap in his knowledge, the anomaly. In how far the anomaly is understood by the user depends on his cognition of the particular situation. Belkin introduced two levels of specificability: the cognitive level and the linguistic level. The cognitive level refers to what degree the user is able to specify (understand) his current situation. The linguistic level refers to the degree the user is able to specify his information need in linguistic terms. Belkin states that if a user is not able to understand his current situation at the cognitive level well enough, then he will hardly be able to express his information need at the linguistic level. He suggests a system design that is built around the user and his ASKs. He refers to Oddy's dialog approach as a good example for such a system design [7] [8].

2.3 Berrypicking approach

In 1989 Bates discovered that the relevant documents are not only the documents which are retrieved at the end of the search, but also some of the documents encountered during the search [9]. He proposed a new approach, which accounts for the changing information need during the search. In every step of the search the user may reformulate his information request based on the knowledge gathered in previous steps. The user is also allowed to keep some of the retrieved documents as relevant. His approach is an evolving search like Oddy's, but differs in that the relevant documents are collected step by step like berries are picked in the forest. Therefore the approach is named berrypicking. In addition he observed that users tend to change their search strategy depending on their rational information need.

2.4 Clustering approach

The above approaches assume some kind of interaction between system and user. In contrast clustering infers from the structure of the document collection on the information needs that could be satisfied with the document collection. Document clustering was subject to research since the 1960s [10] [11] [12]. In 1979 van Rijsbergen formally connected clustering and information need by formulating the cluster hypothesis, which states that closely associated documents are relevant to the same information request [11]. Therefore clustering algorithms highlight patterns in a document collection and allow the users to browse for the needed information. The explosion of digital stored information during the 1990s made this approach very attractive. However, many design questions are still open, most namely the evaluation of document cluster quality [13] [14].

3 Essentials of Information Need Assessment

Based on the historic review in the previous section the following definitions intend to clarify the concept of information need.

Definition 1 (Information Need). *Information need refers to the amount of all absence information, which is necessary for a user to reach his or her goals in a particular situation. The following assumptions hold:*

1. *The user may not know what exactly his information need is.*
2. *The user may not be able to formulate his information need.*
3. *The information need of a particular user may shift during a search session.*

Definition 2 (Rational Information Need and Radical Information Need). *Let $I(U, S)$ be the information need of user U in situation S. The part of the information need the user is aware of is referred to as rational information need I_{Rt}. The part of the information need the user is not aware of is referred to as radical information need I_{Rd}. Rational and Radical information need are disjunct:*

1. $I_{Rt}(U, S) \cup I_{Rd}(U, S) = I(U, S)$.
2. $I_{Rt}(U, S) \cap I_{Rd}(U, S) = \emptyset$.

Definition 3 (Information Need Assessment). *Information need assessment refers to the process of increasing the degree of rational information need of a user during a search session.*

4 IR Assessment Model

The IN Assessment approaches are not competing with each other for which one is the best. Instead each approach fits a certain search context better than the others. IR system interfaces should account for this and dynamically adapt to the user's search context. The model in Figure 1 shows the IR Multi-Interface Model, which incorporates different IN assessment approaches.

The model consists of three layers built around the user. The inner layer represents the interfaces. Every interface gives the user another view on the data. The middle layer represents the engines, which are necessary to realize the interfaces. The outer layer represents the coordination system. The coordination system decides what interface is presented to the user in a particular situation.

For the coordination system to work the classification framework in figure 2 is applied. The framework classifies IN assessment methods along two dimensions: the assessment time and the assessment style.

The assessment time refers to the timeframe in which information is gathered about the user. In the case that the system encounters an unknown user, who demands just in time information, the assessment time is short-term. This situation is common for mass-user internet search engines. In the case that the system continuously collects data about the information need of its users, the

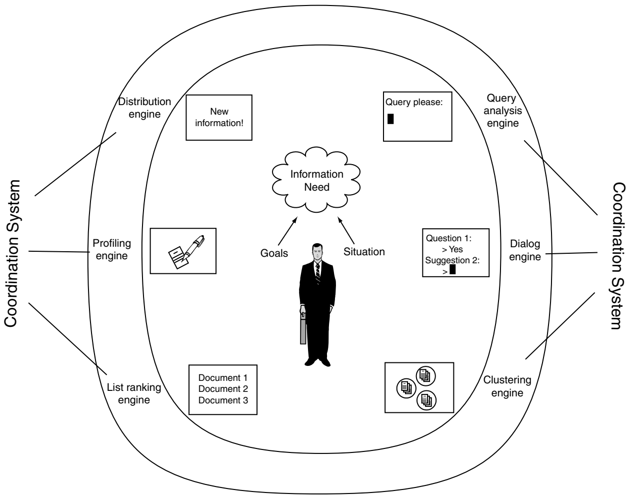


Fig. 1. Multi-Interface Model: The IR system is build around the user. It offers different interfaces for searching in the system's database.

assessment time is long-term. The advantage of long-term IN assessment is that the system can distribute new relevant information to its users when it enters the system. However, for this setting the users should have, at least to some degree, a constant information need over the time.

The assessment style refers to the degree of human/computer involvement in the IN assessment process. If the user formulates his information need by himself, then the assessment style is supervised. This style is very useful when the user knows what source he is looking for. If the system assesses the information need of the user, then the assessment style is unsupervised. This situation is very common when a user acquaint himself with some new topic and does not know the important keywords. But also in the case that an overwhelming amount of relevant information exists unsupervised methods are useful to discover some structure in the information. If both, the user and the system, are involved in the IN assessment, then the assessment style is semi-supervised.

The assessment style is closely tied to the degree of rational IN/radical IN. The higher the degree of rational information need in relation to radical information need the more likely a supervised method will support the user and vice versa. Therefore a search usually starts with an unsupervised or semi-supervised IN assessment method and moves during the search session torwards a supervised method.

5 AiSearch

AiSearch is a Web document retrieval engine developed by the Knowledge-based Systems Group at Paderborn University [3]. The engine is used for research in information retrieval. For the purpose of information need assessment the engine

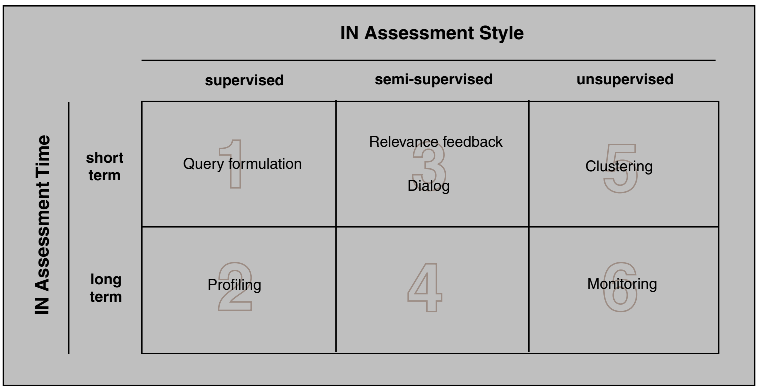


Fig. 2. IN Assessment classification: The IN assessment approaches are classified along the two dimensions assessment style and assessment time. The transparent numbers indicate the degree of IR system involvement in the IN assessment. They range from one (low IR system involvement) to six (high IR system involvement).

incorporates different user interfaces. Up to now two clustering based interfaces are implemented and a third, which highlights chronological relations between documents, is subject to research.

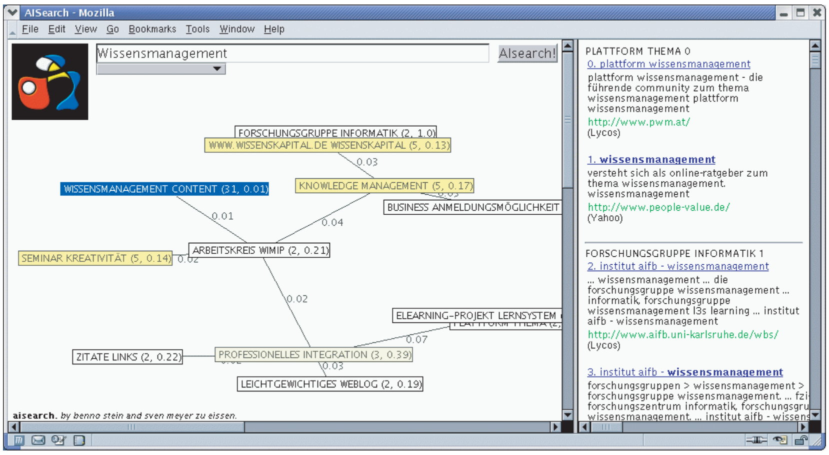
5.1 Implemented Interfaces

Figure 3 shows a clustering based IR interface. In this view the retrieved documents are clustered into conceptionally similar groups. The groups are represented by rectangulars and their content is described by terms on the corresponding rectangular. The content of the selected cluster, which is always centered, is shown in the window on the right side of the screen. The conceptional distance between two clusters are indicated by the distribution of the rectangulars on the screen. Therefore the closer a cluster is located to the center the more closely it is related to the selected cluster. In addition a numbered line between two clusters indicate their closeness in quantitative terms.

In contrast the screenshot in Figure 4 shows a taxanomic view of document clusters. In this view only a small number of all clusters are displayed. The clusters are represented by a term, which describes the content. When a user clicks on one of the terms the corresponding cluster is extended and the view displays its subtopics. The view is very useful when the information need is highly unspecific and the IR system returns a large number of different clusters. In this case a presentation of all clusters at the same time would confuse the user.

5.2 Future work

An interface that highlights chronological relations between documents is subject to current reseach. The basic idea is that knowledge about the development of a


AISearch - Mozilla
File Edit View Go Bookmarks Tools Window Help
AIsearch!
FORSCHUNGSGRUPPE INFORMATIK (2, 1.0)
WWW. WISSENSKAPITAL.DE WISSENSKAPITAL (5, 0.13)
0.03
WISSENSMANAGEMENT CONTENT (31, 0.01)
KNOWLEDGE MANAGEMENT (5, 0.17)
BUSINESS ANMELDUNGSMÖGLICHKEIT
0.01
0.04
SEMINAR KREATIVITÄT (5, 0.14)
0.02
ARBEITSKREIS WIMIP (2, 0.21)
0.02
ELEARNING-PROJEKT LERNSYSTEM
0.07
ZITATE LINKS (2, 0.22)
PROFESSIONELLES INTEGRATION (3, 0.39)
0.03
LEICHTGEWICHTIGES WEBLOG (2, 0.19)
aisearch. by benno stein and sven meyer zu eissen.

Wissensmanagement

PLATTFORM THEMA 0

0. plattform wissensmanagement

plattform wissensmanagement - die führende community zum thema wissensmanagement plattform wissensmanagement

http://www.pwm.at/

(Lycos)

1. **wissensmanagement**

versteht sich als online-ratgeber zum thema wissensmanagement. wissensmanagement

http://www.people-value.de/

(Yahoo)

FORSCHUNGSGRUPPE INFORMATIK 1

2. institut aifb - wissensmanagement

... wissensmanagement ... die forschungsgruppe wissensmanagement ... informatik, forschungsgruppe wissensmanagement l3s learning ... institut aifb - wissensmanagement

http://www.aifb.uni-karlsruhe.de/wbs/

(Lycos)

3. institut aifb - **wissensmanagement**

forschungsgruppen > wissensmanagement > forschungsgruppe wissensmanagement. ... fzi-forschungszentrum informatik, forschungsgru wissensmanagement. ... institut aifb - wissens

Fig. 3. Cluster-based view: The documents are clustered in conceptionally similar groups. The rectangulars represent the clusters, the terms on every rectangular describe the content, and the line between two connected rectangulars indicate their closeness. On the right side of the sceen the content of the selected cluster is displayed in ranked order.

certain topic over time is useful in some situations. Figure 5 shows schematically two views on chronological structured documents. The view on the left side shows a visualization for clusterd results. The vertical axis represents the clusters and the horizontal axis the timeline. Circles in the coordination system represent documents. The bigger a circle the more documents of the corresponding topic refer to events at that time. The view on the rights side shows the "chronological environment" of the current document.

The realization of the engine for the chronological analysis demanded the construction of a knowledge base. At the core of the knowledge base is a set of manual tagged text documents. The tag structure is used to extract time/event entities. A time/event entity is for example the sentence "He plans to change to another club in 2005.". It is called time/event entity because the sentence describes an event that takes place at a certain time. Every single time/event entity is used as an example in the knowledge base database. Figure 6 shows a screenshot of the engines rule manager and a set of examples. The structure of every example is finegrained with additional tags like ¡Year¿ and ¡/Year¿ or ¡Number¿ and ¡/Number¿. Based on the examples and a set of principles the system automatically indentifies time/event entities in texts.

At the moment the engine is still a prototype and its result quality subject to current research. A more detailed description of the system and its performance in practical settings will occur in follow-up publications during this and next year. In addition the content of the texts is restricted to sports topics. However an extension to political and business topics is planned.

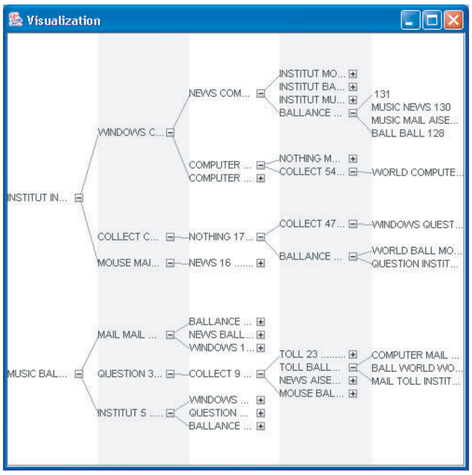


Fig. 4. Taxonomic view: A small number of all cluster is displayed at the beginning. Every cluster is represented by a term, which describes its content. The user can extend the clusters to display subtopics.

6 Summary and Outlook

The purpose of this paper was to shift the eye of the reader to the importance of information need assessment. Therefore the text started by criticising the shortcomings of current IN assessment practices, namely the query input/list output IR systems. A historical survey showed that a user is embedded in different search contexts, which determine how much the user knows about his current information need. The IR Multi-Interface Model was presented to address the existence of several search contexts and it was stated that an IR system should offer different user interfaces and views on the data. Finally the search engine AiSearch was surveyed to demonstrate the functioning of different interfaces in practice.

For the furture the Knowledge-based Systems Group at Paderborn University plans to introduce more interfaces for AISearch. In the short run the view on chronological structured documents will be added to the system and performance statistics will be published in follow-up papers.

References

1. Oddy, R.: Information retrieval through man-machine dialogue. Journal of Documentation **33** (1977) 1–14
2. Belkin, N.: Anomalous states of knowledge as a basis for information retrieval. Canadian Journal of Information Science **5** (1980) 133–143
3. Stein, B., zu Eißen, S.M.: Aisearch: Category formation of web search results. Technical report, Paderborn University (2003)
4. Luhn, H.: The automatic creation of literature abstracts. IBM Journal of Research and Development **2** (1958) 159–165

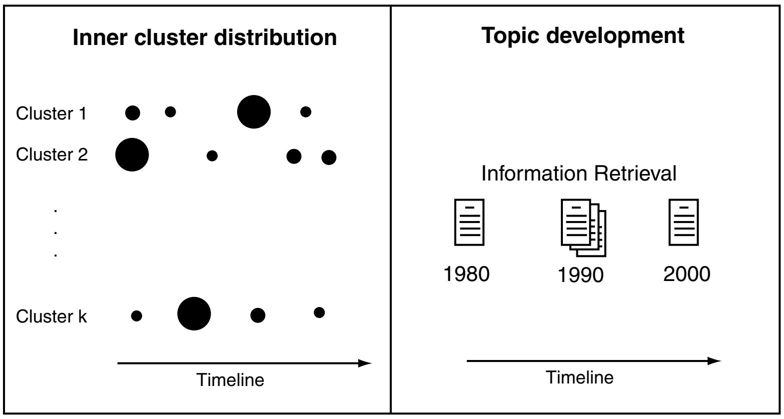
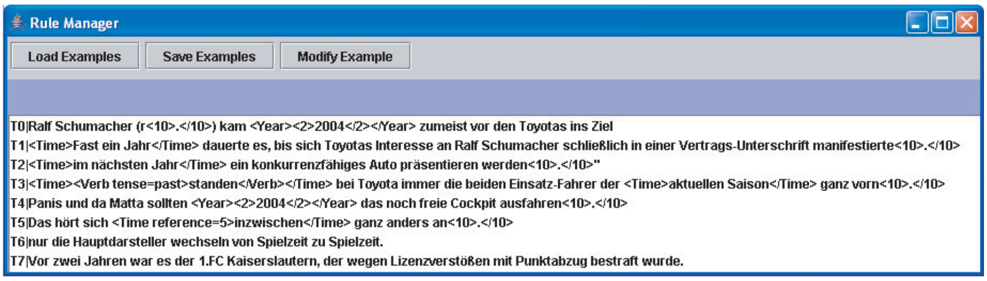


Fig. 5. Views on chronological structured documents.



```
T0|Ralf Schumacher (r<10>.</10>) kam <Year><2>2004</2></Year> zumeist vor den Toyotas ins Ziel
T1|<Time>Fast ein Jahr</Time> dauerte es, bis sich Toyotas Interesse an Ralf Schumacher schließlich in einer Vertrags-Unterschrift manifestierte<10>.</10>
T2|<Time>im nächsten Jahr</Time> ein konkurrenzfähiges Auto präsentieren werden<10>.</10>"
T3|<Time><Verb tense=past>standen</Verb></Time> bei Toyota immer die beiden Einsatz-Fahrer der <Time>aktuellen Saison</Time> ganz vorn<10>.</10>
T4|Panis und da Matta sollten <Year><2>2004</2></Year> das noch freie Cockpit ausfahren<10>.</10>
T5|Das hört sich <Time reference=5>inzwischen</Time> ganz anders an<10>.</10>
T6|nur die Hauptdarsteller wechseln von Spielzeit zu Spielzeit.
T7|Vor zwei Jahren war es der 1.FC Kaiserslautern, der wegen Lizenzverstößen mit Punktabzug bestraft wurde.
```

Fig. 6. Screenshot of the rule manager of the chronological analysis engine.

5. Salton, G., Lesk, M.: The smart automatic document retrieval system - an illustration. Communications of the ACM **8** (1965) 391–398
6. Rocchio, J., Salton, G.: Information optimization and interactive retrieval techniques. In: Proceedings of the AFIPS-Fall Joint Computer Conference, Part I. Volume 27. (1965) 293–305
7. Belkin, N., Oddy, R., Brooks, H.: Ask for information retrieval: Part i. Journal of Documentation **38** (1982) 61–71
8. Belkin, N., Oddy, R., Brooks, H.: Ask for information retrieval: Part ii. Journal of Documentation **38** (1982) 145–164
9. Bates, M.: The design of browsing and berrypicking techniques for the online search interface. Online Review **13** (1989) 407–424
10. Doyle, L.: Semantic road maps for literature searchers. Journal of the ACM **8** (1961) 553–578
11. Rijsbergen, C.: Information Retrieval. Buttersworth, London (1979)
12. Salton, G.: Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley (1988)
13. Stein, B., Meyer zu Eißen, S., Wißbrock, F.: On Cluster Validity and the Information Need of Users. In Hanza, M., ed.: Proceedings of the 3rd IASTED

International Conference on Artificial Intelligence and Applications (AIA 03), Benalmdena, Spain, Anaheim, Calgary, Zurich, ACTA Press (2003) 216–221

14. Stein, B., Meyer zu Eißen, S.: Automatic Document Categorization: Interpreting the Perfomance of Clustering Algorithms. In Gnter, A., Kruse, R., Neumann, B., eds.: KI 2003: Advances in Artificial Intelligence. Volume 2821 LNAI of Lecture Notes in Artificial Intelligence., Springer (2003) 254–266