Evaluating the Intelligibility of Medical Ontological Terms

Björn Forcher[1], Kinga Schumacher[1], Michael Sintek[1], and
Thomas Roth-Berghofer[1,2]

[1] Knowledge Management Department,
German Research Center for Artificial Intelligence (DFKI) GmbH
Trippstadter Straße 122, 67663 Kaiserslautern, Germany

[2] Knowledge-Based Systems Group, Department of Computer Science,
University of Kaiserslautern, P.O. Box 3049, 67653 Kaiserslautern

{firstname.lastname}@dfki.de

Abstract. The research project MEDICO aims at developing an intelligent, robust and scalable semantic search engine for medical documents. The search engine of the MEDICO demonstrator RadSem is based on formal ontologies and is designated for different kinds of users, such as medical doctors, medical IT professionals, patients, and policy makers. Since semantic search results are not always self-explanatory, explanations are necessary to support requirements of different user groups. For this reason, an explanation facility is integrated into RadSem employing the same ontologies for explanation generation. In this work, we present a user experiment that evaluates the intelligibility of labels provided by the used ontologies with respect to different user groups. We discuss the results for refining our current approach for explanation generation in order to provide understandable justifications of semantic search results. Here, we focus on medical experts and laymen, respectively, using semantic networks as form of depiction.[3]

Key words: justification, graphical explanation, semantic search, evaluation, medical terms

1 Introduction

The research project MEDICO aims (among other things) at developing an intelligent semantic search engine for medical documents and addresses different kinds of users, such as medical doctors, medical IT professionals, patients, and policy makers. The ultimate goal of the project [1] is to realize a cross-lingual and modality-independent search for medical documents, such as medical images, clinical findings or reports. Representational constructs of formal ontologies

[3] This research work has been supported in part by the research program THESEUS in the MEDICO project, funded by the German Federal Ministry of Economics and Technology (grant number 01MQ07016). The responsibility for this publication lies with the authors.

are used to annotate and retrieve medical documents. Currently, the MEDICO demonstrator RadSem [2] employs the Foundational Model of Anatomy (FMA) [3] and the International Classification of Diseases, Version 10 (ICD-10)[4]. As there is no existing ontology of the ICD-10 available we implemented a tool which parses the English and German online version providing an OWL ontology of the ICD-10.

Since semantic search results are not always self-explanatory, explanations are helpful to support users who have various intensions to use the search engine. Each user group has different requirements and comes with different a priori knowledge in the medical domain. Medical IT professionals, for instance, may want to test the search engine. In this context, explanations are interesting when the system presents unexpected results. It may turn out that the implementation or the used ontologies are incorrect. Hence, explanations can help to correct the system or to improve it. In contrast to medical IT professionals, patients and citizens are not interested in the exact implementation of the search algorithm. Instead, they may want to learn something about the medical domain. This concerns first of all medical terms but also the connection between medical concepts.

For addressing these issues, we integrated an explanation facility into RadSem. The facility is used to justify search results by revealing a connection between search and annotation concepts. Finding a connection the facility also exploits the mentioned ontologies. Thus, the connection or justification contains further concepts of the FMA or ICD-10. Especially the FMA provides several medical terms for labeling a specific concept. As medical laymen cannot associate any label with corresponding concepts a justification may not be understandable to all of them. In contrast, medical experts may prefer explanations that fit their daily language. In other words, the problem is to select appropriate labels with respect to different user groups. For this reason, we conduct an experiment and discuss its results in order to refine the explanation generation specifically to medical experts and laymen.

This paper is structured as follows. The next section gives a short overview about relevant research on explanations. Section 3 presents current techniques of semantic search algorithms and motivates the need for explanations. Section 4 contains our work of justifying semantic search results. Section 5 describes the user experiment and discusses its results in order to realize a tool that can be used to tailor explanations to different user groups. We conclude the paper with a brief summary and outlook.

2 Related Work

The notion of explanation has several aspects when used in daily life [4]. For instance, explanations are used to describe the causality of events or the semantics of concepts. Explanations help correcting mistakes or serve as justifications.

[4] `http://www.who.int/classifications/apps/icd/icd10online`

Explanations in computer science were introduced in the first generation of Expert Systems (ES). They were recognized as a key feature explaining solutions and reasoning processes, especially in the domain of medical expert systems such as MYCIN [5].

Explanation facilities were an important component supporting the user's needs and decisions [6]. In those early systems, explanations were often nothing more than (badly) paraphrased rules that lacked important aspects or too much information was given at once [7]. For that reason, Swartout and Moore formulated five desiderata for ES explanations [8] which also apply for knowledge-based systems, among them *Fidelity* and *Understandability*.

Fidelity means that the explanation must be an accurate representation of what the ES really does. Hence, explanations have to build on the same knowledge the system uses for its reasoning. Understandability comprises various factors such as *User-Sensitivity* and *Feedback*. User-Sensitivity addresses the user's goals and preferences but also his knowledge with respect to the system and the corresponding domain. Feedback is very important because users do not necessarily understand a given explanation. The system should offer certain kinds of dialog so that users can become clear on parts they do not understand.

In [9], the Reconstructive Explainer is presented producing reconstructive explanations for ES. It transforms a trace, *i. e.*, a line of reasoning, into a plausible explanation story, *i. e.*, a line of explanation. The transformation is an active, complex problem-solving process using additional domain knowledge. The degree of coupling between the trace and the explanation is controlled by a filter which can be set to one of four states regulating the transparency of the filter. The more information of the trace is let through the filter, the more closely the line of explanation follows the line of reasoning. This approach enables a disengagement of an explanation component in order to reuse it in other ES. We took up this theme in our current work.

The Semantic Web community also addresses the issue of explainability. The Inference Web effort [10] realizes an explanation infrastructure for complex Semantic Web applications. Inference Web includes the *Proof Markup Language* for capturing explanation information. It offers constructs to represent where information came from (provenance)or how it was manipulated (justifications). Inference Web includes different tools and services in order to manipulate and present the explanation information. The goal of our research is also to provide tools and algorithms using formal knowledge such as ontologies for explanation provision. The focus of our work is to generate understandable and adequate explanations for knowledge-based systems.

3 Semantic Search

There are diverse definitions of the term *semantic search*. In general, search processes comprise three steps, *i. e.*, query construction, core search process, and visualization of results [11]. In this work, we refer to the most common definition and use the term *semantic search* when formal semantics are used during any part of the search process [12]. In this context, two main categories of semantic search can be identified: fact and semantic document retrieval. Fact retrieval engines are employed to retrieve facts (triples in the Semantic Web) from knowledge bases based on formal ontologies. Such approaches apply three kinds of core search techniques: *reasoning*, *triple based*, *i. e.*, structural interpretation of the query guided by semantic relations, and *graph traversal search* [12]. Semantic document retrieval engines search for documents which are enriched with semantic information. They use additional knowledge to find relevant documents by augmenting traditional keyword search with semantic techniques. Such engines use various *thesauri* for query expansion and/or apply *graph traversal* algorithms to available ontologies [12, 13]. Analogously, the same semantic techniques are used to retrieve other kinds of resources, *e. g.*, images, videos, where additional formal knowledge is used to describe them.

The MEDICO Demonstrator RadSem uses formal ontologies to annotate medical documents in order to describe their content. The search algorithm exploits the class structure of these ontologies to retrieve documents that are annotated with semantically similar concepts with respect to a certain search concept. For instance, searching for radiographs of the hand, users may obtain documents that are annotated with the concept *index finger* or *pisiform bone.* Currently, RadSem employs the FMA and ICD-10 ontology.

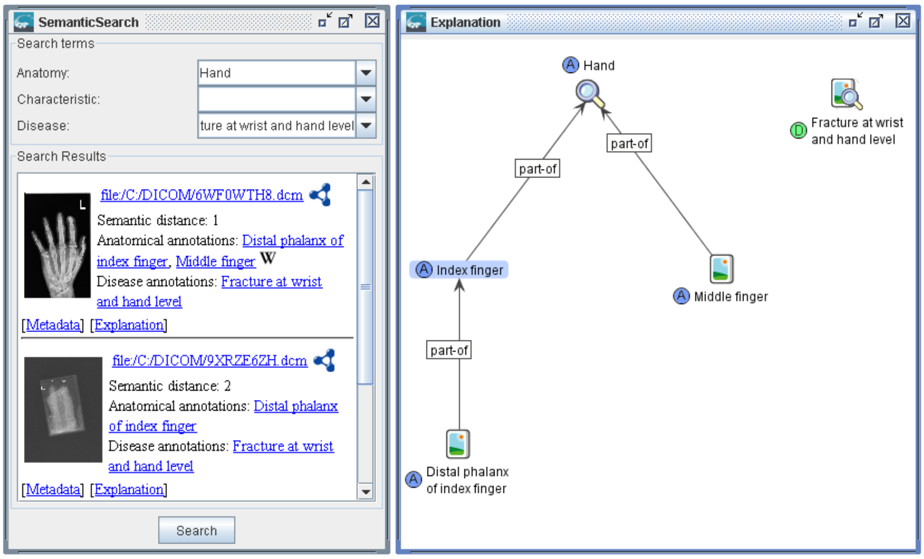
Users have various intensions to use semantic search engines. For instance, a user wants to inform himself of a medical concept he do not remember. In this case, he most probably searches for are similar or superior concept. Imaging, the user searches for information about the *shoulder height* but using the term *shoulder* for his search. If the user obtains a document and associated text snippet highlighting the term *acromion* he may not know whether the document is relevant or not. In this context, a short explanation can provide useful information to support the user's search intention. An explanation expressing that the term acromion is a synonym for shoulder height and that the shoulder height is part of the shoulder may help the user to remember.

The explanation has to reveal the connection between the query and the obtained document. In general, users are not interested in the search techniques of the engine, *i. e.*, how the document is retrieved. In daily tasks users require only a simple justification of the result. As semantic search algorithms use semantic techniques such as ontologies this formal knowledge can be leveraged to generate appropriate explanations.

4 Explanations in RadSem

The explanation facility in RadSem comprises two components: the *Justification Component* and the *Exploration Component.* As its name implies, the first component is primarily intended to justify the retrieval of medical documents. The other component can be used to explore the underlying ontologies and offers various kinds of interaction.

In general, explanations (like any kind of knowledge) have two different aspects: form and content [14]. Explanations are communicated through a certain form of depiction such as text or semantic networks [15]. With respect to the *Understandability* desideratum we chose semantic networks because they are an intelligible alternative to text [16] representing qualitative connections between concepts.



SemanticSearch

Search terms

Anatomy: Hand

Characteristic:

Disease: ture at wrist and hand level

Search Results

file:/C:/DICOM/6WF0WTH8.dcm

Semantic distance: 1

Anatomical annotations: Distal phalanx of index finger, Middle finger

Disease annotations: Fracture at wrist and hand level

[Metadata] [Explanation]

file:/C:/DICOM/9XRZE6ZH.dcm

Semantic distance: 2

Anatomical annotations: Distal phalanx of index finger

Disease annotations: Fracture at wrist and hand level

[Metadata] [Explanation]

Fig. 1. Justification in RadSem

Most probably, a detailed explanation of the search algorithm used in RadSem is not important for most MEDICO users. Reusing our approach in other semantic search projects we ignore consciously the desideratum *Fidelity*. Hence, the Justification Component performs a kind of reconstructive explanation as described in Section 2 omitting all process information of the search algorithm. In this case the search concepts correspond to the input and the annotation concepts correspond to output in the line of explanation, whereas the story in between is constructed by the explanation facility using the ontologies FMA and ICD-10 as knowledge base.

Since search and annotation concepts belong to ontologies the construction is very simple. In general, ontologies can be transformed into a semantic network representing a mathematical graph. Thus, the construction of the line of explanation for semantic search in MEDICO can be reduced to a shortest path problem. We chose the Dijkstra Algorithm [17] to solve this problem. The algorithm can only be performed on non-negative edge path costs, so the question which costs to choose for properties of the ontologies arises. In our first implementation we assume an equal distribution, *i. e.*, all properties have the same cost. Figure 1 depicts an example search in RadSem and according justification.

This simple approach already reveals two general problems. The first issue is with the generation itself. The Dijkstra Algorithm determines only one shortest path. Hence, potential alternative explanations are not found which may be better in a certain context with respect to different user groups. In addition, the number of concepts and thus, the amount of information is preset. Potentially, the explanation path contains too much or too few information. The second problem concerns the adequacy of a justification. In particular the FMA provides several synonyms to label a concept. Currently, the explanation facility uses the *preferred label* to name a certain concept in the explanation path. Most probably, not all users can associate the preferred label with a corresponding concept.

Intelligibility is an important aspect of the quality of an explanation and mainly applies for medical layman. In contrast, medical experts may prefer terms which they use in their daily work. For instance, the term *shoulder girdle* may be better for laymen in contrast to *pectoral girdle* which is more appropriate for experts. To conclude, the difficulty is to determine the best label for different user groups such as medical experts and medical laymen. In this work, we focus on the second problem. Our goal is provide a simple approach to evaluate labels with respect to the different user groups. This approach may be extended not only to evaluate single labels but also to evaluate alternative explanation paths or justifications.

Beyond question, the degree of knowledge about medical terms has a significant effect on adequacy and intelligibility. Hence, a method is required to determine the degree of knowledge of different user groups with respect to the terms or labels used in an explanation path. Therefore, an inherent constant must be considered.

An intuitive assumption is that the degree of knowledge can be correlated with the frequency of terms in natural language. A useful statistical measure are frequency classes. According to [18], the frequency class of a term t is defined as follows: Let C be a text corpus and let *f(t)* denote the frequency of a term $t \in C$. The frequency class $c(t)$ of a term $t \in C$ is defined as $\lfloor log_2(f(t^*)/f(t)) \rfloor$, where t^* denotes the most frequently used term in C. In many English corpora, t^* denotes the term *the* that corresponds to frequency class 0. Thus, a more uncommonly used term has a higher frequency class. In the following, we refer to any frequency class $c(t) = i$ as c_i.

5 Experiment

We assume that the degree of knowledge of medical terms can be correlated with frequency classes. The more often a term is used in natural language the more users know about that term. In order to verify the applicability of this assumption we conducted a user experiment. In this experiment the test persons should estimate their knowledge concerning several medical terms.

5.1 Experiment Setup

For evaluating the personal estimation of medical knowledge 200 medical terms of the FMA and ICD-10 were selected consisting of one or two tokens. As German is the mother tongue of the test persons, only German terms were considered in order to avoid a distortion of the evaluation with respect to language problems. We randomly selected ten terms for each frequency class $c_{10}, ..., c_{13}$ and 15 terms for each frequency class $c_{14}, ..., c_{21}$. The frequency classes were determined with the help of a service of the University of Leipzig.[5] The first group of terms contains well known terms such as *Schulter* (shoulder), *Grippe* (influenza), or *Zeigefinger* (index finger), which all test persons typically know. In contrast, the second group contains terms that are typically unknown to medical laymen. In addition, we randomly selected 40 terms of the FMA and ICD-10 where a frequency class could not be determined in order to have a greater probability that at least some terms were unknown to medical experts. We refer to the corresponding frequency class as c_{22}.

Table 1. Personal Knowledge Estimation

(1)	the term is completely unknown;
(2)	the term has been heard of, but cannot be properly integrated into a medical context;
(3)	the meaning of the term is known or it can be derived. In addition, the term can be vaguely integrated into a medical context;
(4)	the meaning of the term is known and it can be associated with further medical terms;
(5)	the term is completely clear and comprehensive knowledge can be associated.

The 200 medical terms were randomly subdivided into four tests each containing a varying number of frequency classes. Every test person was allowed to do only one test. Thus, we had to take care that each of the four tests was done as often as any other one. All test persons had to estimate their knowledge about each term of a test on a scale from 1 to 5 (see Table 1) indicating their *Personal Knowledge Estimation* (PKE).

5.2 Evaluation

In total, thirty-six persons participated in the experiment: twenty-eight laymen and eight medical experts. The two groups were differentiated as follows. Test persons with a profound medical qualification were classified as experts. For instance, this concerns medical staff, students and doctors. All other test persons were classified as laymen. Figures 2 and 3 depict the result of the evaluation.

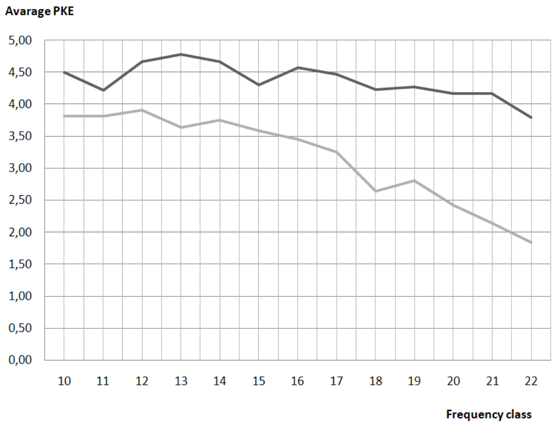


Fig. 2. Experiment results: average values for experts (black) and laymen (gray)

Figure 2 depicts an average value of the PKE as function of the frequency classes for experts (upper curve) and laymen (lower curve) as well. Figure 3 depicts the corresponding standard deviation.

Figure 2 contains two outliers for medical laymen: c_{13} and c_{19}. The first one can be traced to the term *Atlas*. It is an ambiguous term whose meaning in a geological context is quite common. In contrast, its meaning as first cervical vertebra is relatively unknown. The second outlier can be traced to some compounds which are quite common for the German language. The meaning of those terms can easily be derived but their occurrence in daily language is rare. In contrast to laymen, the curve of medical experts is without any irregularity. Merely the estimation of general terms is very interesting because experts seem to consider what they do not know with respect to the general term. The standard deviation for both groups is quite interesting. From frequency class c_{18} on the values jump up. A possible reason for that may be that people have more knowledge in subfields of the medical domain than in others, *i. e.*, when they have a certain disease.

The main objective of the experiment was not to verify a correlation between users' degree of knowledge and frequency classes. In fact, the intention is primarily to denote intervals of frequency classes as a means of prognosis whether user groups probably know a term or require supporting information. For this purpose, we introduce three Boolean functions: $k(t)$ for known terms, $s(t)$ for support requiring terms, and $u(t)$ for unknown terms. With respect to average PKE of medical laymen, we identified three suitable intervals, and defined the functions as follows (index l indicates laymen):

1. $k_l(t)$ is true iff $c(t) \in [c_{11}, ..., c_{15}]$
2. $s_l(t)$ is true iff $c(t) \in [c_{16}, ..., c_{19}]$

[5] `http://wortschatz.uni-leipzig.de/`

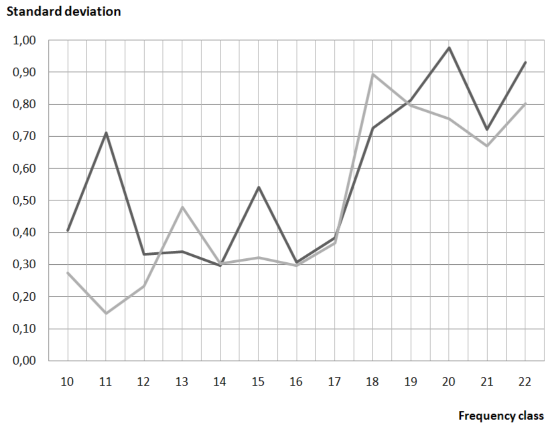


Fig. 3. Experiment results: standard deviation for experts (black) and laymen (gray)

3. $u_l(t)$ is true iff $c(t) \in [c_{20}, ..., c_n]$ and $n > 20$

The proposed functions do not apply to medical experts. The average PKE of all concepts indicates that medical experts generally know terms used in the FMA and ICD-10. Thus, only the function $k_e(t)$ can be defined which is always true (index e indicates experts).

As mentioned before, the functions allow evaluating a justification as presented in Section 4. For instance, there are two justifications A and B of the same search result whereas both comprise three terms. If the middle term of A is a *known term* and if the middle term of B is a *support requiring term*, probably justification A is the better one. The concepts can also be used to tailor justifications. Let a justification represent a path in class hierarchy and comprise four terms. If one of the mid terms is an *unknown term* and the one is *known*, the unknown term can be removed.

In many cases, labels of the FMA or ICD-10 contain other concept labels. For instance, *distal phalanx of left index finger* includes the concept labels *distal phalanx*, *left* and *index finger*. All labels have different frequency classes and thus, a prediction whether a user knows such a concept cannot be made (this applies for all non lexical labels). But this may not be necessary in order to select the most suitable label for a concept with respect to medical laymen or experts. Using the $k_l(t)$ and $k_e(t)$ it is possible to define two sets of labels. These sets can be generalized with respect to various attributes of the labels such as average frequency class of sub labels, label length or token count. The most prominent member of one class can be used to solve a label selection problem. A label with minimal distance to that member may be the most appropriate label for a concept concerning different user groups.

The presented experiment and proposed method may only be seen as a first approach to improve the current explanation generation. We ignored some important aspects of the experiment, such as compounds or ambiguous terms. In addition, users probably may not estimate their knowledge hundred per cent correctly. For this reason, the presented approach can only be regarded as initial point to evaluate terms or complete explanation paths which can be improved by using further methods such as user interactions.

6 Summary and Outlook

In this paper we presented the explanation facility of the MEDICO Demonstrator RadSem. The semantic search engine of RadSem uses formal ontologies to annotate and retrieve medical documents. The explanation facility employs the same ontologies and uses reconstructive explanations as a means of justifying semantic search results. Improving the justifications we conduct an experiment with medical experts and laymen. The objective of the experiment was to determine a correlation between users' degree of knowledge and medical terms. We discussed the results and proposed a method which can be used to determine which terms of the used ontologies should be used in an explanation with respect to medical experts and laymen as initial start. The overall approach can be used to justify various semantic search algorithms using formal ontologies.

The next step of our research is to refine the method for tailoring and evaluating terms and explanations. In addition, we will consider various kinds of user interactions to improve this method.

References

1. Möller, M., Sintek, M.: A scalable architecture for cross-modal semantic annotation and retrieval. In Dengel, A.R., Berns, K., Breuel, T.M., eds.: KI 2008: Advances in Artificial Intelligence, Springer (2008)
2. Möller, M., Regel, S., Sintek, M.: RadSem: Semantic annotation and retrieval for medical images. In: Proc. of The 6th Annual European Semantic Web Conference (ESWC2009). (2009)
3. Rosse, C., Mejino, J.L.V.: The Foundational Model of Anatomy Ontology. In: Anatomy Ontologies for Bioinformatics: Principles and Practice. Volume 6. Springer (2007) 59–117
4. Passmore, J.: Explanation in Everyday Life, in Science, and in History. In: History and Theory, Vol. 2, No. 2. Blackwell Publishing for Wesleyan University (1962) 105–123
5. Buchanan, B.G., Shortliffe, E.H., eds.: Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Addison-Wesley Publishing Company, Reading, Massachusetts (1984)
6. Swartout, W.R., Paris, C., Moore, J.D.: Explanations in knowledge systems: Design for explainable expert systems. IEEE Expert **6**(3) (1991) 58–64
7. Richards, D.: Knowledge-based system explanation: The ripple-down rules alternative. In: Knowledge and Information Systems. Volume 5. (2003) 2–25
8. Swartout, W.R., Moore, J.D.: Explanation in second generation expert systems. Second generation expert systems (1993) 543–585
9. Wick, M.R., Thompson, W.B.: Reconstructive expert system explanation. Artif. Intell. **54**(1-2) (1992) 33–70

10. McGuinness, D.L., Ding, L., Glass, A., Chang, C., Zeng, H., Furtado, V.: Explanation interfaces for the semantic web: Issues and models. In: Proceedings of the 3rd International Semantic Web User Interaction Workshop (SWUI'06). (2006)
11. Organizers, T.: Guidelines for the trecvid 2007 evaluation (2007)
12. Hildebrand, M., Ossenbruggen, J., van Hardman, L.: An analysis of search-based user interaction on the semantic web. Report, CWI, Amsterdam, Holland (2007)
13. Mäkelä, E.: Survey of semantic search research. In: Proceedings of the Seminar on Knowledge Management on the Semantic Web. (2005)
14. Kemp, E.A.: Communicating with a knowledge-based system. In Brezillon, P., ed.: Improving the Use of Knowledge-Based Systems with Explanation. (1992)
15. Ballstaedt, S.P.: Wissensvermittlung. Beltz Psychologische Verlags Union (1997)
16. Wright, P., Reid, F.: Written information: Some alternatives to prose for expressing the outcomes of complex contingencies. Journal of Applied Psychology **57 (2)** (1973) 160–166
17. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische Mathematik **1** (1959) 269–271
18. zu Eissen, S.M., Stein, B.: Intrinsic plagiarism detection. In: ECIR. (2006) 565–569