The Representation of Knowledge Contained in Technical Documents: the Example of FAQs (Frequently Asked Questions)

Evelyne Mounier, Céline Paganelli

▶ **To cite this version:**

Evelyne Mounier, Céline Paganelli. The Representation of Knowledge Contained in Technical Documents: the Example of FAQs (Frequently Asked Questions). Eighth International ISKO Conference, Jul 2004, London, United Kingdom. pp.287-291. hal-00975633

HAL Id: hal-00975633
https://hal.science/hal-00975633

Submitted on 8 Apr 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Evelyne Mounier, Céline Paganelli
Cristal-Gresec, Université Stendhal – Grenoble, France

The Representation of Knowledge Contained in Technical Documents: the Example of FAQs (Frequently Asked Questions)

Abstract: This article deals with the representation of knowledge contained in FAQs. Various works showed that it is conceivable to categorise the various information units contained in technical documents according to the type of the information conveyed. Such a model based on various types of information units makes it possible to represent the knowledge contained in technical documents. Besides, this model proposes a method for the automatic recognition of the information units types contained in these documents. This model has been constructed for "traditional" technical documents and it has been validated with expert users of these documents. In this paper, we propose to validate and extend this model to FAQs by an experimental study with a group of users, expert of the technical field described in FAQs.

1- Context

The present study deals with expert users in a technical domain using an information retrieval system because they need information to perform a task.
A way of considering the indexing of such documents is to study the requests the users address to these documents and to study the information units, which constitute responses for these users. Various works showed that it is conceivable to categorise the various information units contained in technical documents according to the type of the information conveyed (Paganelli, 2002).

In the same way, it was shown that the requests of expert users could be categorised according to the type of expected information.
Thus, we have shown that when expert users search for information either in order to obtain the description of a task to be executed (these requests belong to PRO type) or in order to obtain a description of an object (they are the requests of the OBJECT type). The OBJECT type is declined in more precise sub-types. (Mounier, 2003) Sub-type DFI relates to the definitions, sub-type DF describes an object by its functions (functional description) and the sub-type DO relates to descriptions of objects with their physical or constituent aspect.
Each one of these types presents specific linguistic markers (Clavier, 1997, Mounier 2003), which can be automatically extracted.

2- Validation and extension of the model to FAQs

In this paper, we propose to extend this typology to the FAQs (frequently asked questions), available on the web.
Strictly speaking FAQs do not exactly constitute technical documents**.** There is a conventional structure to FAQs as they are constituted of questions followed by answers but answers and questions are not structured in any particular way, they are not subjected to any rule of form, neither of presentation, nor of structuring.
Nevertheless, FAQs seem to be relatively close to the "traditional" technical documents we have studied up to now. First, like the "traditional" technical documents (maintenance handbook, instruction manual...), FAQs are related to a very specialised field. Thus, one will find FAQs about Unix, XML, Latex, electricity and magnetism... Second, FAQs apply to a particular type of identified public: expert or novice users. Finally, FAQs intend to technical information retrieval.
Moreover, FAQs seem all the more interesting as they present differences compared to technical documents. First of all, and contrary to traditional documents, FAQs record at the same time the answers and the questions (Crowston, 1999).
Besides, FAQs are built from the information needs the users express. Thus, they can enable us to constitute an important corpus of users' requests.

Then, we know that in traditional technical documents, the answers are information units, which must be extracted automatically from the document, whereas in the FAQs, the answers are built specifically to answer the asked questions. Thus, contrary to technical documents, in FAQs, each answer is autonomous and does not belong to a whole document.
This study should make it possible to check if these differences influence the linguistic characteristics of the information units.

3- Methodology

Our work is based on an experimental study. A corpus of FAQs, which apply to expert users, is collected by a search with the Google engine. These FAQs relate to the systems Unix, SQL and Linux. For each system, the first ten results were retained, excluding the FAQ written with an educational aim (for example those dealing with tutorials of programming) and the FAQs that explicitly apply to a novice public. This corpus, made up of 504 answers and 423 questions, has been categorised with the typology previously applied to technical documents.
For the experiment, a selection of units (questions and responses) has been submited to the subjects: 25% of PRO, DO and DF types, and 12,5% of DFI and IND types composed our experimental corpus.
Questions and answers were separated. In general, each answer is made up of a paragraph but answers of several paragraphs can also be found. In this case, answers are not cut down. Each corpus of questions and answers was submitted to a group of expert subjects of the field described in the FAQ. The subjects were asked to categorise each answer and each question with simple values: PRO (descriptions of tasks) or OBJECT (descriptions of objects), and for the units classified with OBJECT: DO (physical or constituent descriptions), DF (functional descriptions), DFI (definitions). An "unspecified" value (IND) is also proposed to the subjects.
The results of this experimentation are analysed in the following way: the frequency of each type and/or sub-type in the questions and in the answers, the proportion of questions and answers unspecified (IND), and the analysis of the units (answers and questions) classified IND.

4- Results

4-1- Distribution of the different types

The table bellow presents the way subjects categorised questions and answers.

Type	Questions Distribution	Answers Distribution
PRO	31.8%	21,2%
DO	28,9%	27,8%
DF	21,9%	24,7%
DFI	10,7%	11,6%
IND	6,7%	14,7%

Table 1: categorisation of questions and of answers

Questions and answers categorised with PRO type present specific features. In PRO type units, the user is, tacitly or explicitly the agent, verbs are mainly in the infinitive or imperative form, the answers are sometimes made up of lists of numbered actions.
Responses and questions categorised with DF type present the following features: the explicit agent is the system or a component of the system, verbs are at the active voice.
In the questions and answers classified as DO: the explicit agent is the system, verbs are stative verbs.
The requests classified as DFI correspond to questions like "*what is X ?*", "*what does Y mean ?"*. In these questions, there is no mention of agent and verbs are mainly stative verbs.
Finally, questions and responses categorised as IND seems to set problems as they can be understood and interpreted in different ways.

4-2- Agreements and Divergences Analysis

There is more agreements between subjects for questions than for responses. Besides, the agreement rates are higher for questions. The agreement rate between subjects goes from 50 to 100% for 83 % of questions, whereas it goes from 50% to 100% for 59 % of answers.
There are two types of questions that set problems to subjects. Several questions do not directly concern the system. For instance, "*do I have to pay fee?*". Other questions can be understood and interpreted in different ways. For instance, "*my*

hard disk IDE is very slow." In this case, either the user wants to know how to make his hard disk faster or he wants to know why his disk is so slow.
Relating to answers for which there is no agreement, subjects hesitate between all the types for 11%, between DO and DF types for 22%, between DO and DFI types for 22% and between PRO and all the other types for 44%.
The answers for which subjects do not agree are often very long. They are composed of several information units which belong to different types. In these cases, subjects hesitate between DO, DF and DFI types.
Some short answers set problems. Either, they direct the user to a website, or they describe tasks the user cannot do.
When subjects hesitate between PRO type and the others, answers always present the following features: the user is the explicit agent but the answer contains modal verbs (*can* for instance) associated with negative, or it contains assumption (*may be, probably)* associated to users actions.

4-3- Discussion

The analysis of several FAQs shows us that it exists different types of FAQs. It seems that some of them are "well writing" FAQs whose structure and organisation are close to traditional technical documentations, whereas other FAQs are closer to questions/answers systems or to discussions whose organisation is similar to verbal exchanges.
This experiment shows that the model we propose to categorize information units applies better to questions than to answers. Besides, this model seems to apply to "well writing" FAQs.
Nevertheless, subjects met difficulties to make the distinction between the answers of DF, DO and DFI types. There's nothing surprising about that when they are extracted from questions / answers systems, as far as this kind of systems, explicitly aimed at helping users, rarely dissociate the description of a system and the description of its functioning.

5- Future Works

The model we propose, based on the categorisation of the information units contained in FAQs, makes it possible to study

a kind of documents available on the Web, to determine the characteristics and the regularities of these documents, and finally to suggest a representation and an automatic recognition of this type of documents.
Nevertheless, in order to fit FAQs that are closer to questions/answers systems, this model needs adaptations. Thus, it could be improved by a deeper analysis of the linguistic features of questions/answers systems structure.

6- References

Clavier, V., Froissart, C., and Paganelli, C. (1997). Objects and Actions: Two concepts of major interest in information retrieval in full-text databases in Workshop on Applications of Natural Language to Information Systems, NLDB' 97 (Vancouver, 1997).

Crowston, K. and M. Williams. (1999). The effects of linking on genres of Web documents. HICSS-99, Kilea, Hawaii, January 1999.

Johannesson E., Wallström C. (1999). Automatic Analysis and Visualization of Stylistic Genres. IRIS22, Keuruu, Finland, August 1999.

Kwasnik, B. H., Crowston, K., Nilan, M. and Roussinov, D. (2001). Identifying Document Genre to Improve Web Search Effectiveness. The Bulletin of the American Society for Information Science and Technology, 27(2), 2001.

Mounier E., Paganelli C. (2003). Extracción y representación de conocimiento contenido en un documento técnico. Actas del IV Colloquio international de sciencas de la documentacion, Salamanca, Mayo 2003, pp651-656.

Paganelli C., Mounier E. (2002). Vers un système de consultation des documents techniques volumineux par des utilisateurs experts: le système Sysrit in IHM et recherche d'information, Paris : Hermès Sciences Publications, 2002, pp195-228.

Paganelli C., Mounier E. (2003). Information Retrieval in Technical Documents : from the User's query to the Information-Unit Tagging. SIGDOC'03, San Francisco, October 2003.