# User Profile Integration in Knowledge Management Systems

J. Duflou[1], K. Hermans[1], B. Vandermeulen[1], B. De Moor[2]
[1]Centre for Industrial Management, Katholieke Universiteit Leuven, Leuven, Belgium
[2]SCD-ESAT, Katholieke Universiteit Leuven, Leuven, Belgium

**Abstract**
This paper describes an approach for automated user profile generation within the framework of knowledge management systems. Based on the recognition of the importance of knowledge management, especially within R&D driven companies, the goals of the research project McKnow at the KULeuven are discussed. One aspect of this project, namely automatic user profile generation, is focussed on.

In this paper, an appropriate structure for user profiles is proposed, as well as methods to construct such user profiles, and necessary tools for this purpose. The results of a series of validation tests are discussed. Finally, possible applications for these user profiles and future research within this project are discussed and conclusions are drawn.

**Keywords**
user profiles, knowledge management, user profile generation

## 1 INTRODUCTION

### 1.1 The need for knowledge management

Nowadays, knowledge is recognised as a factor that is the key to economic competitiveness for innovative enterprises. There is a general recognition that the value of complex products resides not in the factories and buildings used for fabrication, but in the minds of people who create them [1].

Learning curves, starting in the educational system, and continuing during the years of professional activity, lead to insights and experience, which often make up the most important assets of a company. Especially for innovation driven companies, whose R&D efforts are crucial to the sustainability of their activities, the importance of knowledge can hardly be overestimated. Their value is largely determined by the know-how of their employees, for whom the organisation of continuous knowledge upgrade initiatives is absolutely mandatory. Although the awareness of the economic relevance of knowledge has drastically increased in recent years, these assets remain hard to organise and control. Moreover, as the markets and the size of innovative companies tend to grow at a higher rate, maintaining and controlling knowledge within a company is certainly not trivial.

The availability of large amounts of documents through intra- and internet has fundamentally changed the information dependent procedures and especially design processes within R&D departments of larger companies. Many activities on various levels of a hi-tech company nowadays crucially depend on efficient retrieval and management of relevant electronic documents. Therefore, a large number of knowledge management oriented initiatives have emerged over the last years.

While large enterprises, such as Alcatel and Siemens have created proprietary solutions, a large number of commercial knowledge management systems (KMSs) have become available over the past few years as well. The core of such KMSs is generally made up of well-structured databases. Relevant information can be contained in a range of document formats. Converting this information into available knowledge, however, requires an intelligent access as well as context sensitive cross-links between such discrete pieces of information. Database management systems allow to provide such an efficient access and relevant cross-references. Anticipating the future use of available information is, however, a difficult task, which often prevents the creation of information structures that remain effective over a longer period. For highly dynamic sectors such rather rigid data structures often prove to lack the required flexibility to be adjusted to the rapidly evolving enterprise activities. An important observation, when evaluating the applicability of this type of KMSs, is the fact that their creation and maintenance consume a lot of human resources. This significantly affects the economic benefits to be expected when investing in such solutions. In the case of distributed knowledge generation, the coordination of initial set-up and maintenance activities is often an obstacle to overcome. Furthermore, such structured information retrieval systems can provide little or no information on the competences and interests of the users of the KMS.

Appropriate knowledge management should cover much more than intelligent information retrieval only. When not only document info is stored, but also information about tacit knowledge available through

company employees, a KMS can be much more powerful in providing users with the knowledge they need. For this purpose the profile of field experts available in the company has to be included in the KMS, covering information on his or her activities and background, experience, points of interest, etc. This focus on the user and their characteristics is clearly not present in most current systems. Capabilities to locate competences within larger enterprises, to identify key persons in specific fields of expertise, to evaluate the relevance of knowledge available within potential sub-contractor or partner companies, are examples of required complementary functionalities.

### 1.2 The McKnow research programme

The considerations formulated above have formed the basis for an increased interest in more powerful concepts for effective knowledge management. Therefore, a research programme was established at the K.U.Leuven, called McKnow. The objective of this research programme is to create an experimental research platform for developing new advanced methodologies and supporting algorithms for knowledge management in support of the next generation of KMSs.

The focus is on the following new functionalities: development of a user characterisation system, a document classification system, automatic document contents retrieval, cluster analysis methods, methods for linking users and documents, intelligent information retrieval (including human resources), methods for dynamic updating of user and document profiles and methods to locate knowledge. Also the identification of appropriate security and privacy protecting technologies and exploration of non-textual information characterisation are looked into. For each of these functions, appropriate underlying methods and supporting algorithms are being developed in the framework of the project. The functional blocks that are emerging from this research programme, are designed in such a way that, on a system level, they can be integrated in the architecture of a new generation of KMSs. Development of a robust system architecture is therefore an important complimentary target of the McKnow intiative. A conceptual design overview is sketched in Figure 1.

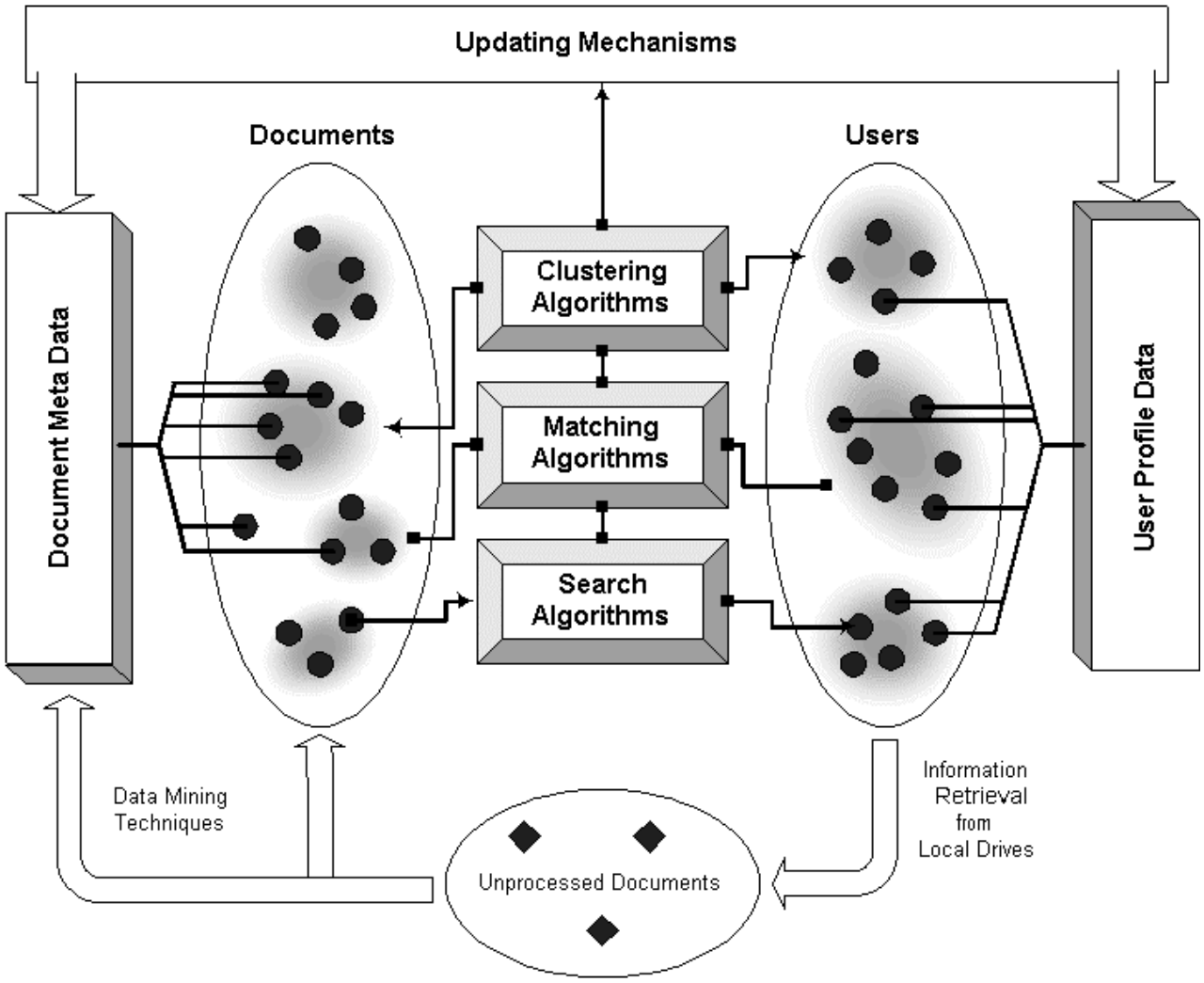


**Figure 1.** Conceptual design overview of the McKnow project

This scheme is characterised by clearly distinguished document and user domains. Extraction of document meta data and automatic generation of user profiles form important modules that provide input for a series of tools based on data-mining techniques. Recognition of clusters in both domains, matching of documents to individual users or clusters of users, intelligent searches for knowledge sources (both document and human resources) are some of the most important functional blocks in this respect. A dynamic feedback system automatically updates document meta-data and user profiles based on the observed user appreciation of consulted documents and the characteristics of newly authored documents.

These new methods, being developed as major scientific objectives of the project, allow information recycling in non-structured environments. Knowledge, captured in different types of documents, and located on accessible network resources, can be analysed, and its relevance for different profiles of users of the system quantified. This approach supports a range of functions for directed information searches. The KMSs based on the McKnow methodology will be able to support a selective information push for newly generated information, taking the profile of the requestor into consideration. Also a dynamic update of user and document profiles is considered, which supports knowledge systems that automatically adjust to the emergence of new fields of interest and an evolving terminology.

The uniqueness of the McKnow research approach lies not as much in the development of efficient browsing techniques for different types of documents, as in the dynamic, bi-directional link between information and automatically updated user profiles. The implications of using such a double, dynamic and automatic characterisation are manifold. Identification of similarities between user profiles allows, for example, more efficient searches for specific information in interactive sessions. Even a search for the most relevant employee expertise, by screening user profiles within larger enterprises, belongs to the possibilities. Another possible application is, for example, risk assessment linked to the departure of field experts through analysing the degree of uniqueness of their user profile.

Methods and supporting algorithms are being developed that allow effective knowledge management in highly dynamic R&D environments. These methodologies cover both intelligent access to document information as well as identification of expertise in human resources.

As a result of the project, the created algorithms and methods can be used as a basis for the development of a new generation of KMSs with the following characteristics:

- User's competences, interests and responsibilities are taken into account through dynamically determined and updated user profiles.
- An intelligent, personalised information access is provided: the dynamic user and document profiles are used for intelligent matching of relevant documents to users.
- Human resources are traceable sources of knowledge.
- Self-learning capabilities automatically improve the effectiveness of the KMS: system utilisation analysis results are used for updating / fine-tuning of both user and document profiles.
- Information from diverse sources and in non-standardised formats can be covered by the system.

## 2 USER PROFILE GENERATION

This paper deals with one specific topic of the McKnow research programme, namely the automated generation of user profiles. In the past few years, user profiles have been studied more extensively, and more and more KMSs aim to integrate user profiles into their system. Different approaches have been described in literature, a overview can be found in [2], and more detailed information about these systems is described in papers [3] [4] [5].

In order to keep the overhead for users as low as possible, we would like to construct a method to initialise user profiles, with some limited user intervention, that could then be automatically updated throughout the process of searching for and looking at information.

### 2.1 Structure of a user profile

In our approach a user is characterised by means of two profiles: a 'knowledge' profile and an 'interest' profile. The knowledge profile is intended to support a human resource competence search, while the interest profile is to be used when people want to search for documents.

As a starting point, each profile consists of a single vector, which is defined in a multi-dimensional vector space where each term represents a dimension and where the values for each dimension are the associated weights for each term [6]. These weights can for example be the number of occurrences of the respective term in the involved documents. The profile vector can be complemented with a region size and a region density.

A region can be seen as the space around a vector in the n-dimensional space, which allows covering for elements that are not present in the vector itself, but are closely related to these topics. This region can be seen as a 'cloud' in the n-dimensional space and allows keeping the number of vectors in a profile relatively low.

The region density is a measure for how dense this region is, in other words for the number of documents that have contributed to this vector in the initialisation phase or during the continuous updating process.

When a user rates documents, the user profile is updated with the document contents. When a document vector is located within the region of a vector in the user profile, this vector will gain weight: the region will not grow, but the density of this region will increase. If the document vector is located outside the region, but still close to it, the document vector can be included and the region can grow. When the document vector to be added is located far from a region in the user profile, a new vector can be established.

By using this approach, a user profile will consist of a (limited) number of vectors, each with a defined region and a region density. Vectors with a small region and a high density are very meaningful vectors, while vectors with a large region and a small density do not allow to provide a very specific expertise description. Also vectors with a small density that have not been changed during a certain period, can be discarded, because they do not represent important user interests or knowledge topics.

### 2.2 Different approaches to generate user profiles

In our project, we consider different approaches to construct a user profile, some of which have been described by other authors and are briefly referred to here for completeness sake.

One possibility is to ask users for a list of terms [7], possibly with a weight associated with each term [8][9]. Although this approach can provide good results, it is usually difficult for a user to objectively characterise himself with a list of terms. Also, such a limited profile may perhaps not be really helpful when searching for documents. However, after updating, when a user has rated a sufficient number of documents implicitly or explicitly, such an initial profile can lead to a good user profile. This kind of profile can also be used as an extra input for a user profile that is constructed with one of the following methods.

To have a better starting profile, we consider another approach: asking the user for a list of a limited number of documents that reflect his expertise or interest. A profile can be created based on the contents of these documents. As a result, a list of terms is obtained, and with each term a number is associated. This number can be the total number of occurrences of a term in all the documents submitted (raw weight), or can be a weighted score, e.g. using tf-idf [10]. Terms that occur frequently in a document (tf = term frequency), but rarely in other documents (df = document frequency), are more likely to be relevant for the characterisation of that document. So the tf-idf weight of a term in a document is the product of the term frequency (tf) with the inverse of the document frequency (idf). Ittner [11] describes an approach to construct a user profile based on the tf-idf vectors of all the documents a person selected. The average of the tf-idf vectors of all interesting documents is used, and a weighted fraction (0.25) of the tf-idf vectors of non-relevant pages is subtracted in order to get a starting vector. This weighing coefficient was determined empirically. Additionally, the user can provide a rating for each document (e.g. 100%, 85%, 90%, … ). Since not all papers will normally be as relevant, a different weight allows taking this into account.

A third approach is based on the job description of a person. A job description for a person is processed like a normal document and a user profile results from this. This last approach has some disadvantages. First of all, a job description is usually rather limited, considering the length of the description as well as the different functions described in it. Secondly, the job description may not be fully representative for the job a person is effectively going to perform: either the job description can already have been out of date by the time a person started working in a specific function, or the content of the job may have shifted since then. Either way, one can see that a job description will not provide good results. A good alternative is, however, to use this job description as an extra input for the user profile, combined with the second approach.

## 3 EXPERIMENTAL VERIFICATION

In order to verify the effectiveness of the proposed approach, a number of experiments were conducted. Five test persons, all members of a single research group and thus somewhat familiar with each others work, were asked to supply us with relevant documents for their research topics (five documents per person). The ranking of terms in the profile vector was determined based on the occurrence of a term in the document collection of that person. Since stemmed words were used, the profiles only contained the ‘stem’ of real words. After the necessary pre-processing steps – like text extraction, language recognition, removal of stopwords and, stemming of the words in a document - the profiles shown in Table 1 were obtained.

When asked to link the test persons to these profiles, people recognised their own profile, and were able to indicate the owners of the other profiles. One error was made on 25 different evaluations (5 people each judged the 5 profiles), as can be concluded from the results listed in Table 2.

As part of the evaluation the similarity between the obtained profile vectors was determined. We can define ‘similarity’ between two user profiles as the

percentage of terms that these profiles have in common. We can write this in a formula:

$$s = \left( \frac{2*c}{a+b} \right) *100\% \qquad (1)$$

where:

*s*: similarity expressed as percentage

*a*: number of terms in user profile A

*b*: number of terms in user profile B

*c*: number of terms that appear in user profile A as well as in user profile B

Note that the applied formula does not take into account the position or the weights of the different terms within a profile. When we calculate the similarities between the different profiles, we obtain the similarity measures in Table 3.

| P 1 | P 2 | P 3 | P 4 | P 5 |
|---|---|---|---|---|
| patient | job | cost | problem | district |
| time | time | pool | rout | problem |
| activ | set | spare | edg | edg |
| product | product | model | product | cycl |
| job | famili | inventori | solut | solut |
| bound | period | airlin | network | opt |
| depart | project | repair | vehicl | optim |
| medic | problem | part | collect | procedur |
| process | process | polici | requir | time |
| servic | date | transship | recoveri | set |
| constraint | number | time | recycl | point |
| hospit | schedule | level | gener | revers |
| network | due | number | model | requir |
| knapsack | earli | system | algorithm | locat |
| analysi | batch | unit | section | search |

**Table 1.** Overview of the generated profiles (only top 15 terms are shown)

| Profile / Person | P 1 | P 2 | P 3 | P 4 | P 5 |
|---|---|---|---|---|---|
| P 1 | P 1 | P 2 | P 3 | P 4 | P 5 |
| P 2 | P 1 | P 2 | P 3 | P 4 | P 5 |
| P 3 | P 1 | P 2 | P 3 | P 5 | P 5 |
| P 4 | P 1 | P 2 | P 3 | P 4 | P 5 |
| P 5 | P 1 | P 2 | P 3 | P 4 | P 5 |

**Table 2.** Assignment of profiles to persons

| Similarity | P 2 | P 3 | P 4 | P 5 |
|---|---|---|---|---|
| P 1 | 27% | 12% | 19% | 31% |
| P 2 | | 19% | 24% | 21% |
| P 3 | | | 20% | 16% |
| P 4 | | | | 44% |

**Table 3.** Similarity measures between the user profiles

Based on these results we can conclude that the profiles of person 4 and person 5 were the most similar pair in the collection. This helps to understand the one error that was made: the assignment of the profiles of person 4 and 5 both to person 5 by one of the participants.

As already mentioned, all five test persons were members of the same research group. When test persons of different working areas are selected, the distinct differences between the respective profiles grow, and user profiles have a lower similarity. Verification of the correct characterisation however becomes more difficult if test persons are not familiar with each other's work.

This approach provides good results. However, the experiment was based on a small sample of documents and of test persons, who are primarily focused on a limited number of clearly related topics. In this context, the approach with only a single vector works fine.

More elaborate tests, taking place in an industrial environment on a larger scale, have been concluded and the first output indicates that this approach delivers user profiles that well characterise the different users.

## 4 POSSIBLE APPLICATIONS

Once user profiles are constructed and have proven their correctness, they can be used for a number of purposes.

One of the most important applications is the matching of documents with user profiles. This can be in a pull situation (match user and document profiles when a user actively performs a search for information) or in a push situation (newly identified documents can be matched against user profiles, offering automatic notification of new relevant information corresponding to one's user profile).

Another possible application is the clustering of users (grouping users together based on the same information needs and interests, with the same background knowledge, with a similar expertise). Since multiple profile vectors are used, users can be a member of more than one cluster. They can shift from one cluster to another one, due to the dynamic updating mechanism for the user profiles.

These clustering techniques can also be useful for the Human Resources Management: by means of user profile clusters the HRM department can get an overview of the various competence domains and reveal certain lacunas in staff competences, which is valuable information for human resource managers.

## 5 CONCLUSIONS

As a result of this research contribution, we would like to come to a situation where well-defined user profiles can be applied with minimal overhead for the users. Since most users are not willing to spend much time on initialising their profile, it is of importance to minimise the required supply of

documents. Initial tests showed that, when R&D staff is asked to submit a relatively small number of documents, possibly combined with some descriptive terms, a good initial user profile can be obtained.

Although our approach provides good results for this small test set, it still needs to be tested on a larger scale to make sure the approach is sufficiently robust, especially when there is more than one vector in a profile. For this purpose, larger scale tests are currently being performed. The resulting user profiles will be tested on their ability to reflect the knowledge and interest topics of the users as well as the ability to select relevant documents.

## 6 ACKNOWLEDGEMENTS

The authors like to recognise the financial support from IWT-Vlaanderen (Instituut voor de Aanmoediging van Innovatie door Wetenschap en Technologie).

## 7 REFERENCES

[1] Tiwana, A., 2002, The Knowledge Management Toolkit – Orchestring IT, Strategy and Knowledge Platforms, Prentice Hall, Upper Saddle River, NJ.

[2] Hermans, K., Duflou. J., De Moor, B., 2003, Automated User Profile Generation For Knowledge Management Systems, Proceedings of the Knowledge Management Aston Conference 2003, 223-233, Birmingham, UK.

[3] Cetintemel, U., Franklin, M.J., Giles, C.L., 2001, Self-Adaptive User Profiles for Large-Scale Data Delivery, Proceedings of the 16th International Conference on Data Engineering, 622-633, San Diego, California.

[4] Korfhage, R.R., 1997, User Profiles and Their Use, Information Storage & Retrieval, 145-161, John Wiley & Sons.

[5] Ackerman, M., Billsus, D., Gaffney, S., Hettich, S., Khoo, G., Kim, D., Klefstad, R., Lowe, C., Ludeman, A., Muramatsu, J., Omori, K., Pazzani , M., Semler, D., Starr, B., Yap, P., 1997, Learning Probabilistic User Profiles: Applications to Finding Interesting Web Sites, Notifying Users of Relevant Changes to Web Pages, and Locating Grant Opportunities, AI Magazine 18/2:47-56.

[6] Berry, M. W., Drmac, Z., Jessup, E.R., 1999, Matrices, Vector Spaces, and Information Retrieval, SIAM Review 41/2:335-362.

[7] Foltz, P.W., Dumais, S.T., 1992, Personalized Information Delivery: An Analysis of Information Filtering Methods, Communications of the ACM 35/12:51-60.

[8] Pazzani, M., Muramatsu, J., Billsus, D., 1996, 'Syskill & Webert': Identifying interesting web sites, Proceedings of the National Conference on Artificial Intelligence, 54-61, Portland, Oregon.

[9] Pazzani, M., Billsus, D., 1997, Learning and Revising User Profiles: The identification of interesting web sites, Machine Learning 27:313-331.

[10] Moens, M.-F., 2000, Automatic indexing and abstracting of document texts, 89-97, Kluwer, Boston MA.

[11] Ittner, D., Lewis, D., Ahn, D., 1995, Text categorization of low quality images, Fourth Annual Symposium on Document Analysis and Information Retrieval, 301-315, Las Vegas, Nevada.

## 8 BIOGRAPHY

Joost Duflou holds master degrees in Architectural Engineering and in Electro-mechanical Engineering. He obtained a PhD in Engineering from K.U.Leuven (Belgium). His principal research activities are situated in the field of design support techniques and management aspects of product development.

Koen Hermans holds degrees in Architectural Engineering and in Industrial Management, both obtained at the KULeuven, Belgium. His current research is focussed around automated and user oriented methods and algorithms for knowledge management.

Bert Vandermeulen holds degrees in Applied Biological Sciences and in Industrial Management, both obtained at the KULeuven, Belgium. His current research is focussed around automated and user oriented methods and algorithms for knowledge management.

Bart De Moor obtained a degree in Electrical Engineering and a PhD in Applied Sciences (Electrical Engineering) at the K.U.Leuven, Belgium. His main research activities are focussed around Numerical Linear Algebra, System Identification and Advanced Process Control, and Neural Networks and Datamining.