Testing Methods and Algorithms for the Next Generation of Knowledge Management Systems

J. Vertommen[1], B. Vandermeulen[1], D. Van Dromme[2], B. De Moor[2], J. Duflou[1]

[1]Centre for Industrial Management, Katholieke Universiteit Leuven, Leuven, Belgium

[2]SCD-ESAT, Katholieke Universiteit Leuven, Leuven, Belgium

Abstract

This paper describes techniques which can be applied in advanced knowledge management systems (KMS) and reports results of two experiments which illustrate the applicability of these methods in real business environments. The techniques are mathematical and originate from the data mining and text mining fields of research, with an added management perspective. The incorporation of user profiles to characterise employees using the system is seen as another typical feature of a next generation of KMS, opening perspectives for better human resource management oriented applications. The need for such systems is motivated and a short description of the used data mining techniques is provided.

The experiments reported in this article demonstrate that there is a good correlation between the clustering results from the developed algorithms and document classifications made by human interpreters. Also, reduction of vector space dimensionality to decrease computation times on high document volumes has shown no problematic loss of classification quality.

Keywords

Knowledge management, KMS, clustering, user profiles.

1 INTRODUCTION

1.1 Need for new KM tools

The fast penetration of digital documents, data carriers and web based communication in today's business has led to new needs concerning the management of these volatile data. Especially the fact that this increasing volume of electronic media also contains a substantial part of the company's valuable knowledge assets has enforced management's interest in good knowledge management tools.

The McKnow[1] project aims at developing advanced algorithms and methods to help dealing with this changed situation in a more efficient way. At the same time, it combines the use of user profiles with a robust approach starting from unstructured data.

1.2 The unconditional automated approach

Contrary to approaches in which documents are classified according to predefined taxonomies and enhanced with metadata, the unconditional approach to data and knowledge management does not impose preconditions about the way the information is initially organised. This has the advantage that it is much less time-consuming because there is no need to build manmade taxonomies or to review and organise the load of digital media.

The following typical example illustrates the usefulness of this approach. When employees leave the company, they often leave a hard drive full of legacy documents which are still valuable, and may even contain the answer to a new employee's questions. The problem is that these media are often hard to access and recycle by other people than the author, because the document owner structured them in a way which seemed most logical to himself, but possibly not to other staff members.

The proposed automatic approach tackles this problem by processing the collection of documents fast and accurately without the need for human review or intervention. Useful data are identified, extracted and transferred to the main knowledge repository, where they stay available to other employees.

1.3 Applications in business

Data mining techniques for automated knowledge identification and extraction are tailored for the World Wide Web, where they are applied to HTML-files from web sites [1]. However, extending the use of similar techniques to documents in a business environment requires a number of adaptations:

- Flexible extraction: the content can be found in a number of different file types.

[1] http://www.mcknow.com/

- Distributed environment: the files are spread over a number of different physical locations, hard drives or network spaces.
- User profiles: the fact that all users of the knowledge system belong to a limited and well-known group of people (the employees of the company) can be used to improve system performance. Integrating a profile for each user creates new opportunities in querying and business analysis. It will not only be possible to link documents, but also to match users with documents and users with other users.
- Security and privacy issues are encountered. Companies do not want to open their databases to third parties, and employees do not like the feeling of being spied on.

2 DATA MINING TECHNIQUES

2.1 Vector representation

Each document or user is represented by a vector, containing stemmed terms and the weights associated with each stem. These vectors (or profiles), the generation of which is briefly explained in Sections 2.2 and 2.3, reside as knowledge entities in an *n*-dimensional space where *n* is the number of unique stemmed terms occurring in the total volume of textual data. With *m* the total number of documents and users, an *n*×*m* matrix allows to represent the knowledge carriers. Typically, such a matrix is very sparse (containing a large number of zeros).

The drawback of such an approach is that it introduces high dimensionality, making it computationally expensive and thus requiring a hardware configuration that is dedicated to this task. Improved algorithms, classification and dimensionality reduction help to overcome this problem.

Cosine proximity[2] of vectors can be interpreted as semantic similarity between the original content of the documents or the user profiles, and forms the basis for clustering.

The cosine proximity between two vectors *r* and *s* is calculated as

$$\cos\theta = \frac{r \cdot s}{\| r \| \| s \|}$$

with θ the angle between vectors *r* and *s* and $r \cdot s$ the standard vector dot product, defined as

$$\sum_{i=1} r_i s_i$$

The norm $\| r \|$ is defined as

$$\sqrt{r \cdot r}$$

[2] Using Euclidean distance in a normalised space will lead to the same result as calculating cosines.

This is illustrated in the experiments reported in Section 3.

By defining different vector spaces, multiple language environments can be covered.

2.2 Document Profile generation

The profile vectors are derived from textual content. First, the language of the document is identified, and then the right stemming algorithm is triggered [2][3]. The stems of the words are counted (indexing) and weighed by a TF-IDF scheme [4]. This row of weighed stems forms the profile vector.

2.3 User Profile generation

The user profiles are generated [5] in a similar way to the document profiles. Currently, each user is represented by a single vector, but it is possible to use a more refined profile consisting of several vectors [6].

One approach for initialising a profile is to collect documents of interest to the user, which can be characterised by means of a document profile as described above. The user profile is then constructed as a combination of the document profiles, respecting the TF-IDF weights in the latter and, if applicable, a ranking of importance between these documents.

2.4 Clustering

Hierarchical clustering is used to increase the speed and accuracy of searching. For this purpose document and user profiles are clustered in a pre-processing step. At first, the search algorithm will allocate a higher priority to the cluster of vectors that is closest to the profile of the user who launches a query.

3 EXPERIMENTS AND RESULTS

3.1 Enterprise context testing

The presented approach was tested with a dataset collected in an international, R&D-oriented company.

3.1.1 Objective

The objective of this test was to prove that the clustering algorithm would group together people working in the same subdivisions of the company, thereby acknowledging a significant similarity between their profiles and a difference with those from employees working in a different subdivision.

3.1.2 Dataset

A set of documents was collected from each of 10 employees at the company. The document-sets varied from 8 to 34 in size. Each of the documents was assigned a weight of importance by the corresponding employee. Besides documents, function descriptions of the test subjects were also collected, from which a part of the company structure could be reconstructed. More precisely, a distinction was made between the R&D department and the IT department.

3.1.3 *Experimental outline*

In a first stage of this experiment, each of the users was assigned a number of profiles, each based on a different amount of documents. A pairwise comparison (by calculating the amount of shared terms) between the profiles in this collection was made per user. From the result of this test it could be concluded that no more than 8 documents were needed to establish a stable profile for a user in this test set.

In a second stage, the user profiles were fed to the hierarchical clustering algorithm, to see if the constructed clusters would reflect the company structure.

3.1.4 *Analysis of results*

The results obtained from the clustering algorithm are shown in Figure 1.

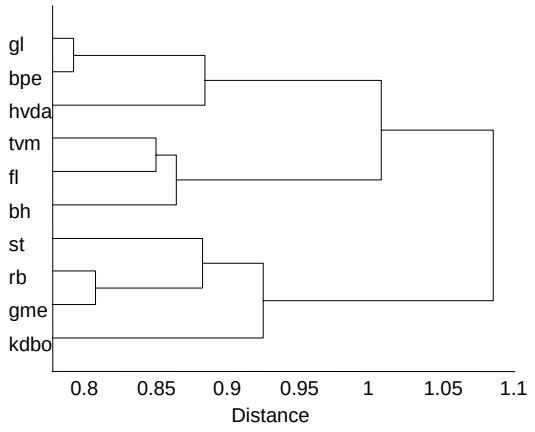

Figure 1. Dendrogram showing the cluster-evolution starting with 10 individual profiles and ending in 1 cluster containing all profiles. The abbreviations in front of each branch are the names of the employees in the dataset. The abcis value is a measure for the distance between clusters.

This dendrogram shows the evolution of the clusters being formed, starting with 10 clusters each containing one profile (at 0.79 or less on the horizontal axis). One by one the profiles are grouped together in clusters, resulting in one big cluster at x-value 1.09. It is clear how, halfway the clustering-process, three distinct clusters are formed (shown in Table 1).

Table 1. The 3 main clusters

Cluster	Members
A	gl, bpe, hvda
B	tvm, fl, bh
C	st, rb, gme, kdbo

When comparing the formed clusters with the company structure, we see that the algorithm has grouped together people who can be expected to have the same interests and/or competences;

Cluster A contains employees that work in the Research and Technology Development division. Cluster C on the other hand contains employees of the IT division. Cluster B required some more thorough analysis since the clustered employees are affiliated with different functional units of the company. When examining the data more in detail it became clear that these were higher-ranked employees and that they were clustered together on the fact that they had more general management interests than specialized technological skills.

This structure is reflected when investigating the most important stems on which profiles are grouped together (See Table 2).

Table 2. Dominant stems per cluster.

Cluster	Stems
A	fast, damp, modal, figur, structural, experimenta, sine, vibraat, …
B	effici, compani, relationship, organisaa, longer, sale, memo,…
C	dataserv, intern, studi, extern, emb, client, xml, securiti, web, intranet

3.1.5 *Summary*

Constructing user profiles based on user-related documents and a TF-IDF weighing-scheme was found to be an effective way of representing a user. This conclusion stems from experiments on the use of the developed clustering algorithm to separate related users from non-related users using the constructed profiles. The experiment reported above forms illustration of this observation.

3.2 Newspaper articles

A second series of experiments was performed on a database of Dutch language newspaper articles. For a number of articles, the author (or KMS user) was known, so these files were used to build the user profiles.

3.2.1 *Objective*

The objective was to test the statement that a user profile based on a limited number of documents is an effective representation of that person's interests and competences.

3.2.2 *Dataset*

The dataset consisted of 1776 newspaper articles (of which the authors were unknown) and 39 articles belonging to 13 known authors, all gathered in April 2003. The authors have different specialisations, as listed in Table 3.

Table 3. Newspaper authors (KMS users) and their topics.

# id	Name (abbr.)	Main topic
1	We.Ma.	Weather forecasting
2	To.Ys.	General domestic news
3	Pe.Va.	Culture (music)
4	Pa.De.	Economics
5	Mi.Do.	International news
6	Ha.Se.	Sports (soccer)
7	Gu.Te.	News & politics
8	Gu.Fr.	General domestic news
9	Fr.Co.	Sports (soccer)
10	Bo.Va.	News & politics
11	Be.Bu.	International news
12	Ba.Do.	Politics
13	Ba.Br.	Politics

3.2.3 Experimental outline

For each author, a vector was created using the textual input from the articles corresponding to those authors, resulting in 13 user profile vectors. For each of the 1776 anonymous articles, a document vector was created.

The vector space containing all user and document profiles (total of 1789 vectors) was normalised and subjected to hierarchical clustering. Results are visualised by a dendrogram as shown in Figure 2.

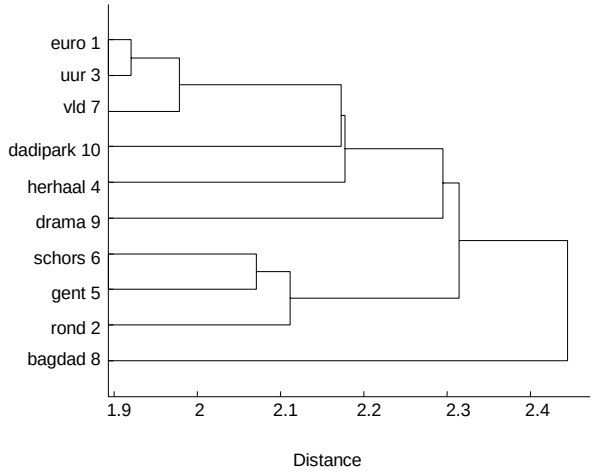

Figure 2. Dendrogram after clustering into 10 clusters. The stemmed terms in front of each branch are the words with the highest overall weight in that cluster, accompanied by the cluster id. The abcis value is a measure for distance between clusters.

The vector space had dimensions 1789×39363 and was processed unmodified in a first stage. Secondly, a reduced space (1789×1000) was created. The differences will be discussed below.

3.2.4 Analysis of results

The coherence of the clusters formed by the algorithm was investigated and it was checked whether the classification can be considered logical from an external observer's point of view.

Looking at the documents in the different clusters, it can be observed that one large cluster and several smaller clusters were identified. The distribution of the documents over ten different clusters is shown in Figure 3.

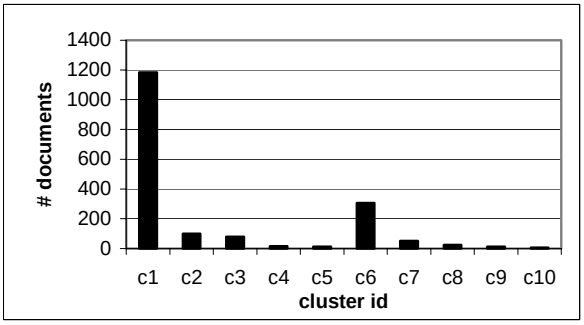

Figure 3. Distribution of documents over 10 clusters.

This phenomenon was merely due to the nature of the dataset, which contained a vast number of often very small news items on various topics. The small clusters are homogeneous and have clearly identifiable topics such as sports (soccer and cycling), movies, or the Iraqi crisis.

Table 4 shows the topics of the clusters and the distribution of the authors over these clusters.

Table 4. Cluster topics. The third column contains the total number of items (documents and users) in the cluster, the fourth shows which users are in each cluster.

Cluster # id	Main topic of cluster	# items	Users in cluster
1	Various news items	1184	1, 2, 3, 4, 5, 8, 11
2	Sports (cycling)	99	-
3	TV programs	80	-
4	Elections	16	-
5	Listings	12	-
6	Sports (soccer)	305	6, 9
7	Politics	50	7, 10, 12, 13
8	Iraqi crisis	24	-
9	Movies	13	-
10	Local news	6	-

This result is satisfying because authors with specific interests are clustered together with documents on the same topic. For example, both authors 6 and 9 have been clustered into the "soccer" cluster, which is where they should be. Also, authors 7, 10, 12 and

13 are grouped together in the cluster about politics, which clearly matches their writings.

An overview of the processing times[3] for the different tasks is given in Table 5. It can be seen that pairwise distance calculation is the bottleneck task.

In a second stage of the experiment, the number of dimensions of the vector space was reduced from 39363 to only 1000 by means of singular value decomposition (SVD) [7][8]. This resulted in a decrease of needed computation time by a factor 211 for the pairwise distance calculation on this dataset.

When comparing the clustering results for this reduced space with the original results, we can only notice a small number of documents (<5%) shifting to other clusters, sometimes even improving the cluster quality. The main structure of the dendrogram remained unchanged.

Table 5. Processing times for the different tasks.

Task	% of total time
Text extraction from original files	0,26%
Indexing and stemming terms	0,40%
TF-IDF weighting	0,04%
Normalisation of vectors	0,50%
Pairwise distance calculation	97,95%
Clustering	0,85%

3.2.5 Summary

The representation of users and documents in a vector space can lead to coherent clusters, based on human semantic interpretation of the texts. This way, users of the KMS can be clustered in a group of documents which are closely related to their area of interest or expertise.

Reduction of dimensionality by SVD of the document-term vector space does not seem to have a negative impact on clustering, so it is safe to apply this technique to reduce calculation time.

4 CONCLUSIONS AND PROSPECT

From the experiments, it can be concluded that the developed profiling and clustering techniques allow to group or differentiate different users according to their fields of interest or expertise. This feature is useful when looking for experts in a specific field of knowledge, when composing teams, or for human resource management in general.

Also, documents can be processed to obtain clusters on the same topic. Moreover, grouping users and documents together creates the possibility to provide users with documents that are of interest to them, which is applicable in an information push system. The experiments indicate that it is possible to achieve an acceptable quality for this task.

Calculation times increase dramatically when the number of knowledge items in the vector space grows, but dimensionality reduction can help to tackle this problem, since it does not seem to have a serious influence on clustering quality.

Further research is oriented towards automatic updating of user profiles based on explicit and/or implicit system feedback. This is expected to result in a dynamic user characterisation that automatically adjusts to evolving user competences and interest domains.

[3] Tasks completed on an Intel® Pentium® III, 1200 MHz with 256 MB RAM, running Windows® XP Professional

5 ACKNOWLEDGEMENTS

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

6 REFERENCES

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

[2] Porter, M.F. (1980) *An algorithm for suffix stripping*. Program, 14(3) pp. 130-137.

[3] Kraaij, W., Pohlmann, R. (1994) *Porter's stemming algorithm for Dutch.* In Informatiewetenschap 1994: Wetenschappelijke bijdragen aan de derde STINFON Conferentie. Noordman, L.G.M. and De Vroomen, W.A.M. (Eds). pp. 167-180.

[4] Salton G., McGill M.J. (1986) *Introduction to Modern Information Retrieval.* McGraw-Hill, Inc., New York, NY.

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

[6] Çetintemel, U., Franklin, M. J. and Giles, C. L. (2001) *Self-Adaptive User Profiles for Large-Scale Data Delivery*. In Proceedings of the 16th International Conference on Data Engineering, San Diego, California, pp. 622-633.

[7] Berry, M.W., Browne, M. (1999) *Understanding Search Engines: Mathematical Modeling and Text Retrieval.* SIAM, Philadelphia, pp. 53-59.

[8] Golub, G.H., and Van Loan, C.F. (1989) *Matrix Computations*, 2nd ed. Baltimore, Johns Hopkins University Press.