Identifying Document Metadata based on multilayer clustering

|  |  |  |
| --- | --- | --- |
| Joris D’hondt[[1]](#footnote-1)[[2]](#footnote-2) | Dennis Vandevenne[[3]](#footnote-3)[[4]](#footnote-4) | Paul-Armand Verhaegen12 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Joris Vertommen | Dirk Cattrysse12 |  | Joost R. Duflou1[[5]](#footnote-5) |  |  |

Abstract

This paper presents a novel technique to semi-automatically identify metadata for documents when installing a knowledge management system. Document management systems often deal with large collections of documents. This vast amount of information needs to be searchable for the knowledge worker. Supporting techniques are needed to aid the knowledge worker in his search for information. Many of these techniques are based on the presence of metadata for each document. The techniques presented in this paper are based on a novel approach called multilayer clustering. Using this clustering technique, documents can be assigned to one or more document types. Besides this assignment to a specific type, properties are assigned to this document. Values for these properties are suggested based on term networks extracted from this document. The preliminary tests presented in this paper were performed on a public and several private dataset. The results obtained from the tests indicate that this approach is promising.

keywords

Knowledge management, metadata, multilayer clustering

1. introduction

The ever-increasing numbers of digital documents available in an industrial environment have boosted the research for techniques to enable the processing of large information repositories. This research led to the notion of Document Management, the capturing and handling of information in company documents. The derived document management systems (DMS) integrate several supporting techniques for the handling of this large amount of information. Search engines and user profiles are some of the attempts to facilitate the retrieval of this information. Many of these techniques rely on the existence of metadata when the documents entered in a DMS.

A large part of the setup of the Document Management System used in this research is, at this moment, performed manually. This manual process contains two steps:

1. identifications of document types
2. assigning labels to each document type

In the DMS an document type is defined as a representative of a group of ‘similar’ documents. Similar documents means either similar in the content they describe or in the layout they contain.

A group of documents that are similar in content are for instance invoice documents. A practical example of documents similar in layout can be a group of documents containing the same template. In Figure 1 a representation is given of this concept. In this document collection three document types are identified. With these document types several documents are linked, represented as empty rectangles.



Figure : Representation of the concept 'Document Type'

A set of labels is assigned to every general document or document type. These labels are the properties of the document type. They describe the document more in detail, for example, an invoice is defined by items such as the company name, account number and project type related to this invoice. These relationships are shown in Figure 2.



Figure : A presentation of the concept 'Labels'

The labels or properties need to be defined for every document type in the DMS.

To perform these identification processes, consultants currently interview several company employees and system administrators to discover the different document structures in the company. During this interview two types of questions are asked to the employee to identify these document types and labels. The first type of questions concerns the view of the employee on the document structure he uses. What kind of documents is he confronted with on a regular basis? What kind of structure or groups would he select? Based on the different opinions, a general structure is conceived containing the different document types. After the identification of the labels, described in the following paragraphs, this structure is used as a starting point for assigning the documents to the document types. A document is at least related to one document type, but if possible, multiple relationships are drawn to document types. These multiple relations are shown in Figure 3, where a document is connected to two types: *Invoice* and *Project*. In this example an invoice is always related to a project, therefore these relationships are identified.

The main idea behind these multiple relations is that the system can create several different viewpoints on the documents present in the system. These different viewpoints deliver a different view on the document structure in the software system. So it is possible to first structure the documents according to the project currently being executed, or to restructure the documents according to clients of the company. These different viewpoints can be associated with the role of an employee in the system.

The document types are not the only items governing these different viewpoints. As already mentioned a set of labels or properties is assigned to every document type. To obtain this set of labels or properties, a second set of questions is asked to the employees such the required fields on a certain type of document. The aggregation of these different answers results in a set of labels for every document type. This list however is not static. In a later stage it is possible to adapt this list to the current needs.

The documents assigned to one or more document types inherit all the labels defined by these related document types. For every label actual filled-in labels needs to be assigned by the user or creator of the document. The complete result of the identification process is given in Figure 3. The already mentioned double relationships with the document types is complemented with actual values for the three identified labels.



Figure : Representation of a document connected to two document types and three metadata

The manual procedure described here needs to be performed once. This procedure however requires hours if not days to obtain several structures in the document collection of a company. This manual work has some serious drawbacks: apart from the time invested by the consultants and the company employees, the final structures depend on the subjective view of the interviewed employees, and also by the number and experience of these employees.

Based on these drawbacks the research was initiated to develop a methodology to shorten and facilitate and automate this identification procedure.

The complete process is characterized as semi-automatic because the system presents the user the possibility to select the document types from a list. The suggested list however contains a ranking.

Summarized the goal of the research presented in this paper is twofold:

1. to semi-automatically identify document types
2. to semi-automatically identify labels and values

2. METHODOLOGY

Figure 4 presents an overview of the followed methodology for identifying a number of document types and a list of possible metadata for each document. In the following sections the different steps of this methodology are explained. The objective of this methodology is to identify document types, and possible properties and values for these properties for each individual document.

The methodology is composed of two steps:

1. Identification of multilayer clusters
2. Term by term distances

The identified multilayer clusters are related with the previously described document types. The multilayer clustering is described in section 2.3.

Every clustering is described by a set of important terms. The distances between every pair of those important terms can give an indication of the content of the document type, and therefore is related to the labels and values of the document types. These term by term distances are described in section 2.4.

2.1. Data conditioning

Before any processing step is executed, a data conditioning step is performed to convert the different document layouts and types to a workable plain text format. This pre-processing step contains, if needed, a conversion of a paper document with OCR techniques to a digital file. Language identification is performed to divide the document collection in single-language document sets. A filtering is also performed to retain only textual documents. In a second phase all layouts of the documents are omitted, and a conversion is made to a general plain text format as input for the next processing steps.

2.2. Vector REPRESENTATION

In order to process a document for obtaining a list of properties and values of the metadata of a document, the document must be further converted into a workable format. A widely used technique is to convert each document into a multidimensional vector, also called the Vector Space Model (VSM) (Salton, Introduction to Modern Information Retrieval, 1986). In this representation format, each dimension of the vector represents a term in the vocabulary of the document. The dimensionality of this vector space is defined by the entire collection of unique terms in the

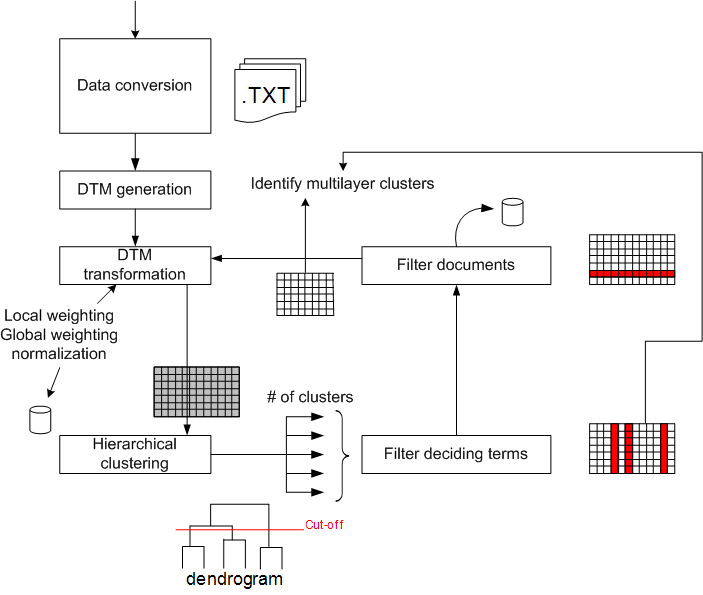


Figure : Overview of the methodology

document set. The values in the vector are the number of occurrences of the related terms. In order to reduce the number of different terms in the vocabulary, different steps are performed to reduce the dimensionality:

1. a stemming operation is performed on each term reducing every term to its root form. The used stemming algorithm is the Porter algorithm (Porter, 1997).
2. Several filtering steps are performed to remove other non-informative words. Stop word, only-number and length filtering are performed to reduce the list.

Not all stems in the document contain a similar amount of information (Newman, 2005). To simulate these differences in informative content, weighting schemes are applied based on certain assumptions to indicate the importance of a stem in a document. In literature a large collection of weighting schemes exist to weight the stems in a document vector (Salton & Buckley, 1988). In this research, the popular TF-IDF weighting scheme (Everitt, Landau, & Leese, 2001) is applied to the stem vectors. To eliminate the bias due to longer documents, a normalization step for every document is performed.

2.3. multilayer clustering

The goal of multilayer clustering is to obtain stable groups of documents. These stable groups are identified by applying multiple clustering procedures and by removing the important stems identified in each clustering steps.

2.3.1. Document by Term Transformation

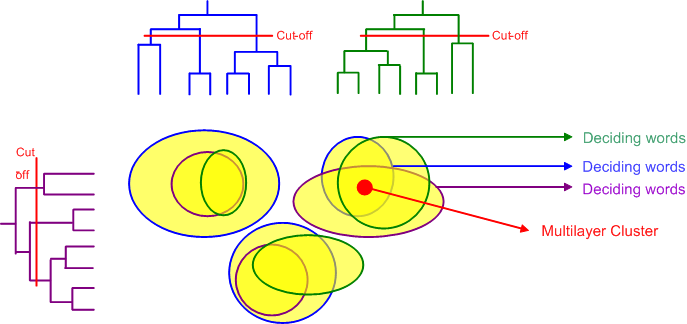
Based on the techniques a sparse document by term matrix can be obtained of the complete document collection. In this matrix, the rows represent the different documents. The matrix representation enables the application of Linear Algebra techniques in this space (Dominich, 2008). The most important of these techniques in the course of this paper, is the definition of a distance measure between vectors, and, consequently, between documents. The dissimilarity measure used in this research is based on an adaptive version of the cosine dissimilarity measure (D'hondt, Vertommen, Verhaegen, Catrysse, & Duflou, 2009).

2.3.2. Hierarchical clustering

Clustering techniques are one of the supporting techniques that enable the user to create an overview in an unstructured collection of documents. In hierarchical clustering (Jain, Murty, & Flynn, 1999) a cluster tree called dendrogram is constructed in the following iterative manner: the closest documents or groups of documents are grouped together. Cutting this dendrogram at a selected level of detail, delivers a clustering output with a corresponding number of clusters. Many clustering techniques assume that the number of clusters is known beforehand. In the methodology presented in this paper, this parameter is predicted with the algorithm described in (Vertommen, Janssens, De Moor, & Duflou, 2008). This algorithm selects different preferred values for the number of clusters, ranked in descending order of likeliness.

Figure : Overview of the multilayer concept

2.3.3. Filtering

Due to the different weighting of the stems of a document, the higher weighted terms will have a higher influence on the clustering result. Once the clustering is determined, these determining stems are filtered from the domain vocabulary. In the document by term matrix, the weights are replaced by zero-values. The clustering step is restarted until a cluster quality criterion is violated. Many criteria, based on compactness and isolation, are available in literature, such as *minimum total distance* or *separate clusters* (Raskutti & Leckie, 1999) If the cluster quality measured with such criteria has dropped below a predefined threshold, the clustering and filtering step is stopped. This parameter influences the granularity of the clustering result (see section 4).

2.3.4. Identification of multilayer clusters

Identifying multilayer clusters involves identifying the stable regions of the multiple cluster results. A stable region of a clustering is defined as documents that tend to cluster together when a clustering is repeated multiple times. When documents are repeatedly clustered together, while iteratively several filtering and clustering steps are performed, an indication is obtained for identifying a new document type for that stable group of documents.

This idea is represented in Table 1. In this table, three multilayer clusters are identified with four repeated clustering. The columns L1 to L4 are the four repeated cluster processes. In these columns, the labels of the different clusters in the different layers are shown. These cluster labels are only important in the same layer, e.g. A1 and A2 are different clusters of a different cluster process. The rows are the three identified multilayer clusters. Groups of documents with the same chain of cluster labels form a stable clustering through successive clusters and therefore can indicate the presence of one or more document types. All documents combined in multilayer cluster MLC1 were grouped together in four repeated clustering’s with the previously described filtering.

Table : Example of three identified multilayer clusters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster ID | L1 | L2 | L3 | L4 |
| MLC1 | A1 | A2 | A3 | A4 |
| MLC2 | B1 | B2 | B3 | B4 |
| MLC3 | A1 | A2 | A3 | C4 |

For instance in Table 1, the documents behind MLC1 and MCL3 are most likely related because they are clustered together in the first three of the four layers, meaning that in each of these layers the important terms defining those clusters were the same. The clustering at every level is thus dominated by a set of important terms. Documents grouped together multiple times can be described by these sets of terms, which therefore contain interesting metadata of the document clustered together. The relationship between these terms is described in the following section.

2.4. Term by Term distance matrix

The identification of multilayer clusters and its related collection of important terms is important for the identification of document types. A second important step is determining the distance between two terms. This distance is defined here as the number of co-occurrences of these two terms in the collection of documents. In other words, this value is equal to the number of times the two terms appear together in the same document. The higher this value of co-occurrence between two terms, the more likely they describe a similar content. In a collection of bills the terms ‘amount’ and ‘account’ will appear frequently together in the document collection. For every pair of terms this ‘distance’ can be calculated and summarized in a term-by-term matrix. The value in every cell is the distance between the respective terms represented by the row and column.

The term-by-term distance matrix is the input for a term graph. The nodes of this graph are the different terms present in the document collection, and the arcs represent the distance between the related terms. This will be explained more in detail in section 3.2.

3. KM functionalities

In this section some of the possible functionalities in a KMS, based on the developed techniques, will be discussed. Some results of preliminary tests will be shown combined with a general description of the desired functionality. The results described here were obtained from two datasets:

1. Newsgroup 20
2. City

The Newsgroup 20 dataset is a collection of approximately 20000 newsgroup articles, equally divided over 20 newsgroup items. Some of the topics are closely related, e.g. pc.hardware and mac.hardware, while others are unrelated, e.g. forsale and religion.christian. An extensive description of the structure can be found in (Lang, 2008).

The community dataset contains around 300 (Dutch) documents collected from the administration of a small city in Belgium. This dataset contains several anonymized documents such as letters, bills and summons. A detailed description of the content is not available.

3.1. Identification of document Types

The identification of document types, based on the previously described techniques, is an important step in de setup phase of the KMS system. Previously this phase was performed manually. A first analysis of the documents selected for the input of the KMS led to the selection of the document types. Interaction with the system administrators and knowledge workers was necessary to achieve this structure.

The development of the described techniques is a first step towards the automation of this identification process. When the KMS is installed on the dedicated server, the application is executed over the selected directories. The result of this application is a cluster structure, governed by the important terms of the clustering as previously described. An example of a result of this multilayer clustering technique on the Newsgroup 20 dataset is shown in Figure 6.

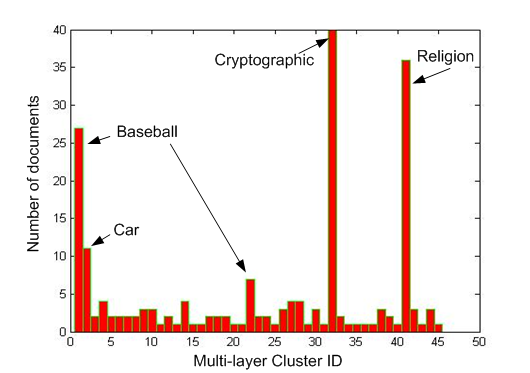


Figure : Multilayer clustering result of the newsgroup test set

From this document collection, 150 documents were selected, originating from four topics. The application of the multilayer clustering resulted in four large document types, corresponding to the selected topics. This is indicated by the F-measure value of 0.73 (Raskutti & Leckie, 1999). Changing the filtering threshold (see section 2.3.3) and the number of iterative clustering steps (see section 2.3.4), changes the granularity of the results. In this figure, the document type ‘Baseball’ is divided in two document types. Reducing the number of iterative clustering steps results in a lower number of document types.

Figure 7 displays the results of the multilayer clustering of the city data set. In this dataset, three document types were identified. These three topics were, as can be expected in this environment, ‘bills’, ‘quotations’ and ‘communications’ to and from citizens. The ‘bills’ and ‘quotations’ document types were strictly delineated. The main reason for these coherent multilayer clusters is the presence of common and limited vocabulary, e.g. due to the usage of templates. The ‘communications’ document type is much less delineated. These communications, inbound and outbound, contain no strict template and important words or phrases such as subject lines, salutations and closing greetings group these documents to a less coherent group.

The functionality and related user interface based on these techniques enables the administrator or user of the KMS to identify the document types present in the system. The semi-automatic aspect in these techniques is that the administrator still has to name the identified document types. This task is facilitated by the ordered set of important terms describing these clusters. Often, the document types are suggested by terms at the top of those lists.

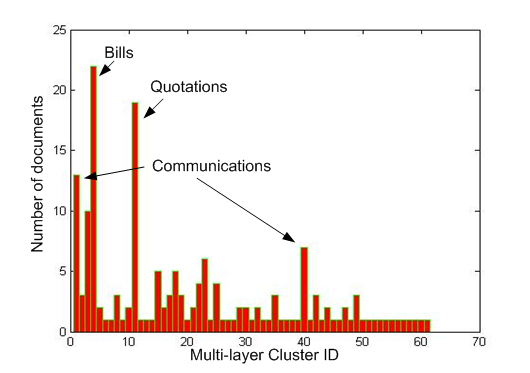


Figure : Multilayer clustering result of the city test set

3.2. Identification of labels and values

As described in the previous sections, every cluster is governed by a set of important terms. After the identification of the types, certain properties or labels need to be assigned to the document types. The process of assigning metadata is aided by the previously described term graph (see section 2.4).

In Figure 8 a term graph is shown, displaying the identified links between the important terms. The terms are grouped according to the quantified distances between the terms. A dense group of terms is therefore related to a document type. However, the terms can be shared across multiple document types.

Based on the terms present in the graph, labels can be identified for the related document types. Figure 6 represents a term graph of the city dataset. For readability reasons this term graph is summarized. No distance values are shown, and isolated terms are omitted. Strongly related terms are grouped together. In this graph the three identified document types can be recognized. The related terms indicate the metadata labels that can be assigned to the document type. E.g. the document type ‘payment’ (number 2 on the figure) can be identified with the terms ‘payment’ and ‘deposit’. The important terms defining this multilayer cluster or the related document type are summarized and translated in Table 2.

Table : English to Dutch translation of terms related to the document type 'payment'

|  |  |
| --- | --- |
| **English** | **Dutch** |
| deposit | storting |
|  | |
| comment | mededeling |
| costumer | opdrachtgever |
| beneficiary | begunstigde |
| receipt | afgifte |
| copy | kopie |
| account | rekening |
| address | adres |
| name | naam |
| letters | letters |
| memo date | memodatum |

This list of terms aids the administrator to semi-automatically assign labels to the document type. The user afterwards can assign a value to every label, depending on the content of the document.

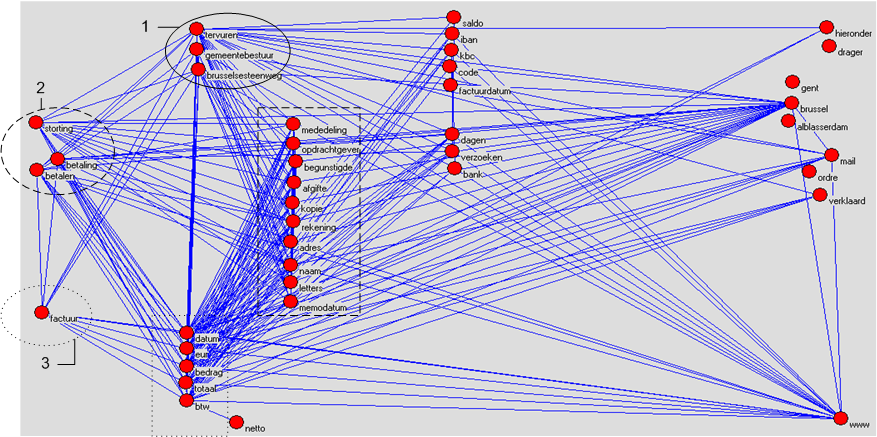
4. Conclusions

Figure : Term graph representation of the city dataset

Metadata is useful information but difficult to gather and maintain. Automated approaches that gather information from the computing environment only cover the easy-to-gather metadata, such as author name and creation date. This information can be obtained from the operating system or the login of the user. Important information only extractable from the content of the document such as the account number or the name of the company in a billing information is more difficult to gather. Depending on statistical information of the terms this paper presents the first steps in a semi-automatic approach to aid the administrator and user in assigning metadata to documents. This process is composed of two steps. Firstly, document types are identified. In this manner every document is assigned to at least one document type. In a second phase a number of metadata labels are assigned to the document types. All documents linked to a certain document type inherit these metadata labels. The whole process has a semi-automatic character: the administrator has to identify the document types from an ordered list of suggestions and has to select the appropriate metadata labels for the document types based on the links in the generated term graph.

The first tests conducted with these techniques indicate that the presented methodology is a promising step to reach the goal of semi-automatically identifying the metadata for documents. The future work includes a refinement in the distance measurement between terms. Also the development of supporting techniques for the user in assigning values for the identified metadata labels still needs to be developed.

5. Acknowledgments

The authors would like to recognize the financial support from IWT-vlaanderen (Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie).

References

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* Wokingham: Addison-Wesley.

D'hondt, J., Vertommen, J., Verhaegen, P.-A., Catrysse, D., & Duflou, J. R. (2009). Pairwise-adaptive dissimilarity measure for document clustering. *Information Sciences* , submitted.

Dominich, S. (2008). *The modern algebra of information retrieval* (Vol. 24). Springer.

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster Analysis.* London: A Hodder Arnold Publication.

Jain, A. K., Murty, M. R., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys* , 264-323.

Lang, K. (2008, 01 14). *The 20 Newsgroups data set*. Opgehaald van http://www.ai.mit.edu/people/jrennie/20Newsgroups/

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* *, 46* (5), 323-351.

Porter, M. F. (1997). An algorithm for suffix stripping. In *Readings in information retrieval* (pp. 313-316). San Francisco: Morgan Kaufmann Publishers Inc.

Raskutti, B., & Leckie, C. (1999). An Evaluation of Criteria for Measuring the Quality of Clusters. *Sixteenth International Joint Conference on Artificial Intelligence* (pp. 905-910). San Francisco: Morgan Kaufmann Publishers.

Salton, G. (1986). *Introduction to Modern Information Retrieval.* New York: Mcgraw Hill Computer Science Series.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal archive* *, 24* (5), 513-523.

Vertommen, J., Janssens, F., De Moor, B., & Duflou, J. R. (2008). Multiple-vector User Profiles in Support of Knowledge Sharing. *Information Sciences* *, 178* (17), 3333-3346.

1. Centre for Industrial Management, Department of Mechanics, Katholieke Universiteit Leuven [↑](#footnote-ref-1)
2. Firstname.lastname@cib.kuleuven.be [↑](#footnote-ref-2)
3. Computer Security and Industrial Cryptography, Department of Electronics, Katholieke Unversiteit Leuven [↑](#footnote-ref-3)
4. Firstname.lastname@esat.kuleuven.be [↑](#footnote-ref-4)
5. Firstname.lastname@mech.kuleuven.be [↑](#footnote-ref-5)