Spring 6-10-2017

LEVERAGING TEXT MINING FOR THE DESIGN OF A LEGAL KNOWLEDGE MANAGEMENT SYSTEM

Jannis Hanke
University of Würzburg, jannis.hanke@uni-wuerzburg.de

Frédéric Thiesse
University of Wuerzburg, frederic.thiesse@uni-wuerzburg.de

LEVERAGING TEXT MINING FOR THE DESIGN OF A LEGAL KNOWLEDGE MANAGEMENT SYSTEM

Research in Progress

Hanke, Jannis, University of Wuerzburg, Wuerzburg, Germany, jannis.hanke@uni-wuerzburg.de

Thiesse, Frédéric, University of Wuerzburg, Wuerzburg, Germany, frederic.thiesse@uni-wuerzburg.de

Abstract

In today's globalized world, companies are faced with numerous and continuously changing legal requirements. To ensure that these companies are compliant with legal regulations, law and consulting firms use open legal data published by governments worldwide. With this data pool growing rapidly, the complexity of legal research is strongly increasing. Despite this fact, only few research papers consider the application of information systems in the legal domain. Against this backdrop, we pro-pose a knowledge management (KM) system that aims at supporting legal research processes. To this end, we leverage the potentials of text mining techniques to extract valuable information from legal documents. This information is stored in a graph database, which enables us to capture the relation-ships between these documents and users of the system. These relationships and the information from the documents are then fed into a recommendation system which aims at facilitating knowledge trans-fer within companies. The prototypical implementation of the proposed KM system is based on 20,000 legal documents and is currently evaluated in cooperation with a Big 4 accounting company.

Keywords: Knowledge Management System, Legal Information Retrieval, Recommendation System, Text Mining.

1 Introduction

Organizations make use of knowledge management systems to leverage their knowledge resources in order to sustain competitive advantage in volatile environments (Kankanhalli et al., 2005). Such constantly changing requirements are particularly evident at the intersection of law and business. The sheer number of laws published on different levels (e.g. international, national, regional) forces enterprises and wealthy individuals to draw upon professional legal services that ensure their compliance with current legal regulations. To keep an eye on changing regulations, law and consulting firms employ open government data published by the public authorities of over 50 countries (Jetzek et al., 2013) and several intergovernmental organizations (e.g. OECD). Especially the vast amount of machine-readable legislative open data serves as a huge and indispensable information repository. However, it also makes finding relevant information a labour and knowledge intensive task, especially since the information is presented differently on each portal (Boella et al., 2012).

Against this backdrop, we propose a knowledge management system that can support and enhance the organizational processes of knowledge creation, storage/retrieval, transfer, and application (Alavi and Leidner, 2001) by legal professions. In more detail, our system addresses the issues at an auditing, tax, and management consultancy, which performs repetitive, time-consuming tasks to group, analyse, and communicate the contents of legal documents. We consider the example of a consulting firm and its group of legal researchers who screen the pages of governments and intergovernmental organizations to identify novel and relevant changes concerning the topics (i) compliance, (ii) taxes, (iii) general legal conditions, and (iv) disclosure. The necessary information within the relevant documents is aggregated manually and evaluated afterwards to create comprehensive reports. By now, the initial documents are not reused or further exploited. We address these existing inefficiencies by proposing an IT solution which automatically collects and stores all kinds of legal and legislative documents. The system enables the retrieval of documents and the knowledge transfer among legal researchers. To this end, we rely on a combination of traditional text mining to extract the most informative parts of legal documents, and recommendation techniques. Our collaborative recommendation procedure is based on the users' history and shares the implicit knowledge about the perceived relevance of documents with co-workers. In addition, the system engages users in explicit knowledge sharing by providing a collaborative editing component. Documents in the database can be labelled manually with key descriptors, which are in turn used to improve the retrieval results of all users.

However, the automatic processing of legal texts by text mining and information retrieval algorithms poses several challenges. Legal texts are formulated in a language which complicates their automatic processing, are multilingual, and require an efficient scalable processing due to the ever-increasing volume of published documents. The present paper addresses these challenges in the following ways. First, we review related literature and highlight gaps of legal knowledge management, which motivate our research. In sum, it can be said, that only few IS studies consider the application of information systems in the legal domain (Knackstedt et al., 2013). Next, we introduce our knowledge management system as an IT artefact applied in the legal domain, following the design-oriented research paradigm according to Hevner et al. (2004) and focus in this research-in-progress on the corresponding design cycle. To conduct a proper evaluation, we develop a web-based user interface in close cooperation with practitioners.

2 Related Work

Alavi and Leidner (2001) describe different views of knowledge which may lead to different perceptions of knowledge management (KM). In the context of our research, we take the perspective that knowledge is a condition of access to information, which implies that KM systems have to provide effective search and retrieval mechanisms for locating relevant information (McQueen, 1998). Plenty of research has been conducted on how to organize and structure documents as a base for knowledge creation in organizations. The information of documents created in the organization was leveraged by Kankanhalli et al. (2011), who investigate the effects of (internal) knowledge reuse through electronic repositories in the context of customer service support. Mourtzis and Doukas (2014) propose to support the design and manufacturing of customised products with a systematic capturing and storing of case-specific knowledge from earlier intragroup engineering projects. However, external sources of information may also be an important mean of knowledge acquisition. Trappey and Trappey (2008) collected patent documents for the implementation of a R&D knowledge management system. Cheng et al. (2009) use news data and industry databases to build a financial KM system. Rodriguez-Enriquez et al. (2016) propose a supply chain KM system to retrieve documents as a decision base for e-procurement by including several external data sources like social media or product websites.

Alavi and Leidner (2001) state that besides creation, storage, and retrieval, the knowledge transfer is an essential domain of knowledge management. This process, through which individuals are affected by the experience of others, can be supported by IT, too (Ko et al., 2005). One possibility are Wikis for the collaborative modification of the underlying content (Wagner, 2004). Several researchers examine the suitability of collaborative document editing for efficient knowledge management (Begoña and Carmen, 2011; He and Yang, 2016; Schaffert, 2006). (Wu et al., 2010) introduce a collaborative authoring approach with similarities to traditional wiki systems. Besides the creation of content, users are encouraged to categorize documents with pre-defined document hierarchies supported by text clustering algorithms. Nevertheless, other researches argue that wikis suffer from several drawbacks (e.g. creation of repositories is laborious and costly, repositories are often ignored by workers) which limit their suitability for knowledge management (Kiniti and Standing, 2013). Therefore, other approaches proliferated in recent years. One promising approach to transfer knowledge among workers are recommendation systems (Ko et al., 2005).

Generally, recommendation approaches can be categorized into (i) content-based methods, where similarity between items (e.g. documents) is taken into account, and (ii) collaborative approaches, where recommendations are based on users with similar preferences (e.g. search history) in the past (Adomavicius and Tuzhilin, 2005). Li et al. (2006), for example, introduce such a collaborative recommendation approach in the context of knowledge sharing on the specific example of music selection. Zhen et al. (2010) propose an inter-enterprise knowledge recommender system to deliver the proper knowledge to the proper people at the example of a manufacturing enterprise in China. The appropriate information is provided by a recommendation engine which is built on top of a rule and constraint set previously assembled by domain experts. To assist users to identify experts in a knowledge management system, Li et al. (2011) suggest an expert recommendation system. The recommendations are generated based on text clustering of previously rated documents in the knowledge repository. Furthermore, Li et al. (2011) argue that using text mining techniques helps to reduce drawbacks stemming from manual assigned categorizations and rules in knowledge management systems.

Unfortunately, in some domains, text mining is especially challenging. Legal and legislative texts pose such an example as they, often contain ambiguous and vague legal terms and furthermore, are typically context- and time-dependent (Knackstedt et al., 2014). For this reason, the automatic processing of legal documents has been extensively researched from a technological point of view. We categorized the corresponding literature based on the utilized text mining techniques and identified three different main groups: (i) classification/clustering of legal texts, (ii) information extraction, and (iii) natural language processing. Francesconi and Passerini (2007), Maat et al. (2010), Lin et al. (2012), Boella et al. (2013), and Lin et al. (2015) try to structure the huge amount of legal and legislative texts by document classification. Information extraction from legal and legislative documents uses predefined rules (Agnoloni and Tiscornia, 2010; Jackson et al., 2003; Varga and Edmonds, 2016) or machine learning (Lippi and Torroni, 2016) in order to resolve references between documents (Tran et al., 2013), or extract key terms (Lagos et al., 2010). Natural language processing techniques allow for building Q&A Systems (Rodrigo et al., 2013), which enables citizens to ask questions about law or implement a topic modelling approaches to monitor emerging topics in legislation (Hagen et al., 2015).

This technology-focused research serves as foundation for comprehensive KM systems in the legal domain. Bianchi et al. (2009) combine several tools to support legal professions in exploring a complex corpus of norms and documents. They propose retrieval techniques in the legal domain and design a user interface, which is then used for the comparison of a corpus of legislative XML-files. Savvas and Bassiliades (2009) address the challenges arising from the great volume of administrative documents. Adopting an ontology, which represents the public administration structure, assists citizens and businesses in interpreting legislative content. Boella et al. (2012) determine requirements for a knowledge management system to maintain regulatory compliance. The authors suggest that the system should consist of a legal ontology to express legal concepts and supporting natural language processing techniques. The challenges are illustrated on a set of European regulations in the financial domain. Other researchers also build their legal knowledge management system on top of legal ontologies (Antonini et al., 2013) and in addition use text mining to classify content automatically (Tello-Leal et al., 2015). Closest to our work is ‘Eunomos’, a legal document management system introduced by Boella et al. (2011). ‘Eunomos’ enables users to research laws and legal concepts via a web-based user interface. It enables traditional keyword retrieval and the search of similar legislations in a database using previously determined topic categories. However, ‘Eunomos’ suffers from a number of drawbacks which motivate our research. Firstly, ‘Eunomos’ classifies documents based on their entire content. We hypothesize that the results of document comparison could be improved by only extracting certain parts (i.e. important key facts) from legal texts. Secondly, Boella et al. (2011) completely neglect aspects that enable the knowledge transfer among users although this is an overall objective of a comprehensive KM solution. We aim to address the challenge of knowledge sharing and transfer by implementing a collaborative recommendation approach. Finally, ‘Eunomos’ stores data in a relational database. We argue, that a graph-based storage is more effective as the main purpose of a recommendation system is to identify relationships between documents and users.

3 Technical Specifications of the Artefact

Figure 1 illustrates the components of our artefact, which are explained in detail in the next sections.

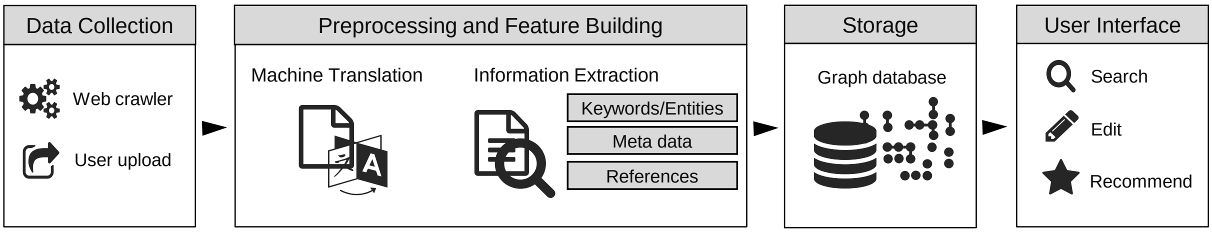


Figure 1. Description of the artefact

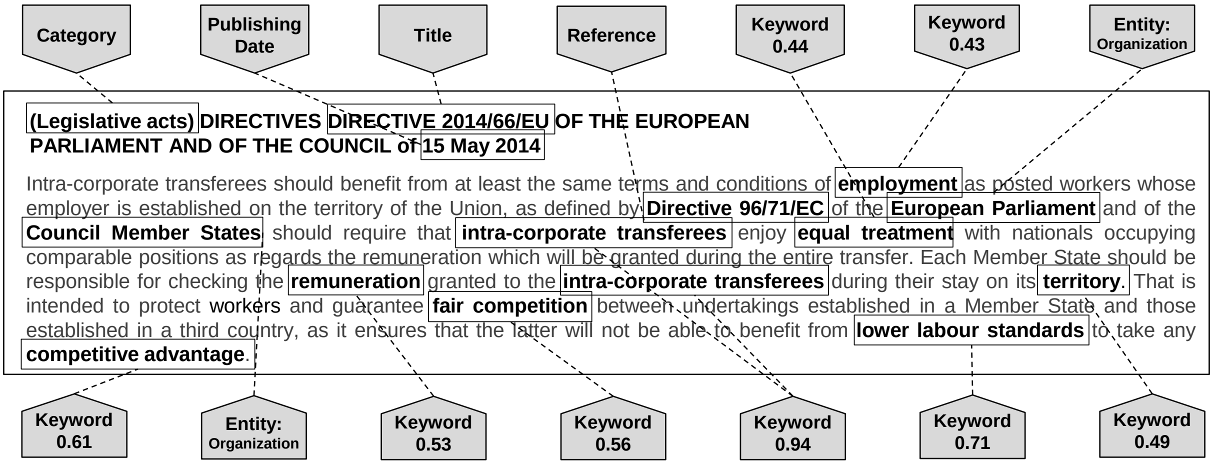
3.1 Data Collection

Our artefact is currently instantiated with a collection of 20,000 legislative texts, which we automatically crawled from government websites. We implemented a system that makes use of a technique called ‘focused web crawling’. Focused crawling is a technique that assesses the relevance of a page before downloading it and therefore enables to collect pages related to a given topic (Diligenti et al., 2000). This is essential, as we only want to collect legal-related publications from the currently fifteen government websites in our repository. For each (differently structured) website we define a set of seed pages as starting point for a recursive document search as described by Tsoi et al. (2003). Furthermore, a predefined set of site-specific rules, like file path, file format (e.g. .pdf), and anchor tags on the HTML page (Pant et al., 2004) are used to determine if a document is relevant before being stored. To keep the database up-to-date, the software automatically revisits each website periodically and downloads novel documents. One drawback is the need for continuous maintaining of the rules if the structure of the target webpages changes. Another source of valuable information are the users of the system themselves. The prototype allows for adding any legal or legislative document to the database that is currently of interest to the user.

3.2 Data Preprocessing and Feature Building

One characteristic of legislative texts is that they are typically written in the official language of the respective country that publishes them. Automatic processing of multi-lingual texts is particularly challenging as text mining is, in simple terms, counting the intersections of words. There are many possibilities to overcome difficulties arising from multiple languages, like for example, Latent Semantic Indexing (Wei et al., 2008) or Artificial Neural Nets (Lauly et al., 2014). These algorithms are usually trained using a parallel corpus (same content in different languages) and able to transfer the learned models onto unseen documents afterwards. Besides such unstructured corpuses, cross-lingual text mining makes use of structured knowledge bases, too. Wikipedia, for instance, is used for document relatedness calculation (Nastase and Strube, 2013; Navigli and Ponzetto, 2012) and text classification (Ni et al., 2011). Another possibility to handle multilingual text is the transformation into a single language using a machine translation approach. This approach has been unpopular in the past due to poor quality, but recent advantages in this field motivated researchers to use machine translation in preprocessing steps (Balahur and Turchi, 2014; Erdmann et al., 2014; Pecina et al., 2014). Our artifact follows a machine translation approach to handle multilingual legal documents and transforms them into a single language. Machine translation may suffer from syntactic problems (e.g. tense or negation) or can have quality issues if the word order in the languages differ. However, as our system relies on individual keywords instead of on the overall textual content, the word order is of minor importance. Another general issue are semantic problems, like polysemy (words with multiple meanings), but with the utilization of deep learning methods, this could be alleviated by considering context windows around each word (Guo et al., 2014). We use the Microsoft Translator API (www.microsoft.com/en-us/translator), which is based on neural networks and hence benefits from the extensive research on word sense disambiguation in recent years (Chen et al., 2014; Gao et al., 2013).

Subsequently, we extract named entities from the documents. Named entities are terms describing real word objects, like for example, companies or locations. (Carreras et al., 2002). While early systems used hand-crafted rules defined by linguists, today the most commonly employed technique is machine learning (Nadeau and Sekine, 2007). These systems are supervised (or semi-supervised) and train themselves on a set of labeled input data (containing positive and negative examples of named entities) in order to automatically induce extraction rules. Nadeau and Sekine (2007) provide a comprehensive literature review on entity extraction and on the frequently employed algorithms. More recent research focuses on the capability of deep learning for entity extraction (Lample et al., 2016; Santos and Guimaraes, 2015). We evaluated different tools and finally choose to apply IBM AlchemyAPI (www.alchemyapi.com), which is based on neural networks and a large proprietary training corpus (Hsu, 2016), as well as a framework called spaCy (www.spacy.io). We consider both tools as black boxes, as they do not provide detailed information about the underlying algorithms. Nevertheless, they have been successfully applied in previous studies (Jiang et al., 2016; Rizzo and Troncy, 2011). We combine both tools by using the overlap of the returned results, similar to Jiang et al. (2016). Thus, we are able to identify five different types of entities: (i) Locations (e.g. countries, regions), (ii) persons, (iii) organizations (e.g. Parliament, foreign affairs council), (iv) field terminologies (e.g. minister, presidential election), and numeric types (e.g. date, monetary values). In addition, we use AlchemyAPI to extract keywords. These keywords (and their importance expressed as percentage value) are also determined using deep learning techniques (Hsu, 2016). The reason for only extracting a limited amount of information instead of exploiting the full content can be attributed to the characteristics of legal language. Knackstedt et al. (2014) describe the language as intentionally vague and ambiguous, which makes the capturing of the relevant parts difficult, especially for computers.



(Legislative acts) DIRECTIVES DIRECTIVE 2014/66/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 15 May 2014

Intra-corporate transferees should benefit from at least the same terms and conditions of **employment** as posted workers whose employer is established on the territory of the Union, as defined by **Directive 96/71/EC** of the **European Parliament** and of the **Council Member States** should require that **intra-corporate transferees** enjoy **equal treatment** with nationals occupying comparable positions as regards the remuneration which will be granted during the entire transfer. Each Member State should be responsible for checking the **remuneration** granted to the **intra-corporate transferees** during their stay on its **territory**. That is intended to protect workers and guarantee **fair competition** between undertakings established in a Member State and those established in a third country, as it ensures that the latter will not be able to benefit from **lower labour standards** to take any **competitive advantage**.

Figure 2. Features extracted from an exemplary document

Normative references found in the majority of legal and law texts are another important feature. Such references indicate that a legal regulation rests on another law, or extends existing regulations. Therefore, we conclude that intersections of references in legal texts indicate that these texts deal with similar topics. Legal references can be extracted by supervised machine learning (Tran et al., 2013) or by predefined rules (Palmirani et al., 2003). We implement an approach with similarities to Palmirani et al. (2003) and define country-specific rules to extract references from law texts. References among legal texts are often differently abbreviated (e.g. TEU for "Treaty on European Union"). To resolve this issue, we collected legal abbreviations from Wikipedia, which offers an extensive list for different languages. Figure 2 shows our extracted features by the example of a legislative act of the European Union. We extracted three types of meta data (category, title, and publishing date), one reference to another EU Directive and nine keywords or named entities. For each keyword, the AlchemyAPI returns a relevance score, which we later use to calculate similarities among documents.

3.3 Similarity Calculation and Recommendation Generation

As our system is mainly focused on the relationship between documents and users, we use a graph-based storage approach. A graph database employs nodes, which are connected by edges to represent the relationship between them. Properties allow to store information that relates to nodes (Angles and Gutierrez, 2008). We store the following elements as nodes: (i) users who interact with the system, (ii) documents (including their full text as property), and (iii) the features we extracted in section 3.2. Documents are connected to users if they assessed them as relevant (via the user interface), and connected to features if they are included therein. Furthermore, the edges between keywords and documents contain the relevance score determined by the AlchemyAPI. Figure 3 illustrates the structure of our graph database.

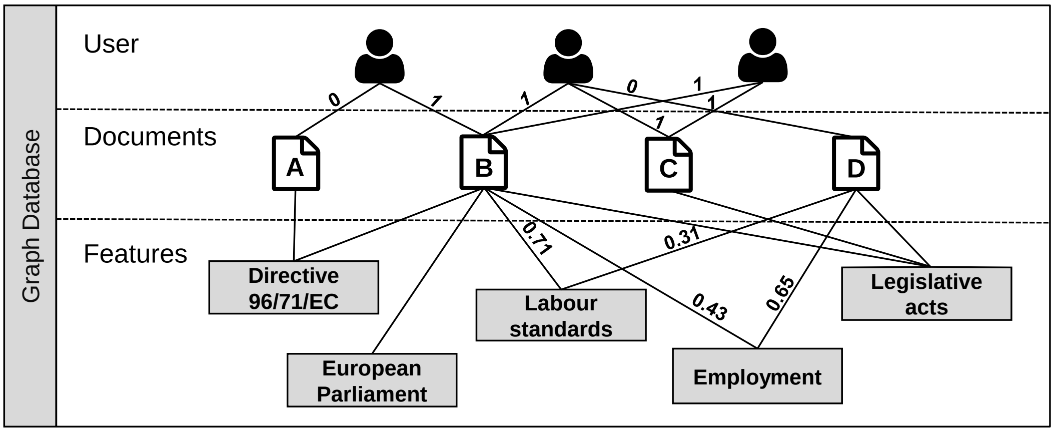


Figure 3. Exemplary excerpt of our graph database

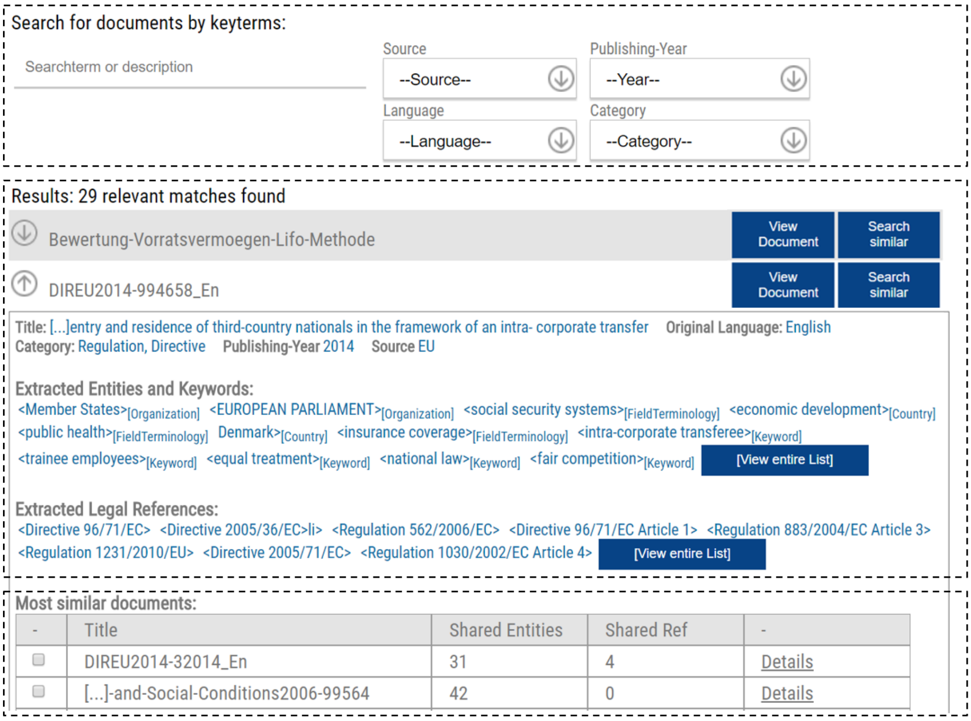
To generate recommendations of potentially relevant documents, a similarity calculation among documents is required. If a document is highly similar to documents a user has considered before, we can assume that this document is of interest to this particular user, too. The graph-based structure allows for the calculation of two different types of similarities: (i) content-based similarity of documents and (ii) document-based collaborative similarity. The rationale behind the content-based similarity is that, similar documents reference to similar features (entities, keywords, references and metadata) in the graph. Therefore, a simple option to determine document similarity is to measure the overlap of common features. However, this may lack of quality since, for example, we want to identify similarities of documents, which do not share any feature in the graph. If we would just use the overlap of features, documents *A* and *D* (see Figure 3) would not have any similarities at all. However, in our graph, these documents are connected by document *B* as *A* and *B* share one feature and *B* and *D* three features. More precisely, *A* and *D* are connected by a short path and, therefore, tend to have a content-related similarity. With an increasing number of nodes and edges, such calculations become computationally expensive. To cope with this problem we implement the SimRank algorithm proposed by Jeh and Widom (2002). SimRank measures similarities of nodes using their relationship to other objects in the graph. A modified version of SimRank allows to take weighted edges and therefore to take the importance of keywords into account (Antonellis et al., 2008).

If a user assesses (via the user interface) a document as relevant (‘liked’) or irrelevant (‘dislike’), the information is stored by creating an edge between the user and the document node labelled with 1 or 0, respectively. Against this backdrop, we calculate an item-based collaborative similarity for documents or, in other words, the similarity between documents using the user preferences in the graph (Deshpande and Karypis, 2004). We assume two documents to be similar if they were often considered relevant or irrelevant together. We employ a modified version of the jaccard similarity coefficient to measure the size of the intersection of ‘likes’ and ‘dislikes’ among all users. The jaccard similarity is one possible measure for recommender systems in case that only binary relevance information is available (IJntema et al., 2010). According to our adopted jaccard coefficient, the similarity between document *A* and *B* is calculated by dividing the total number of users that ‘like’ both documents and the total number of users that ‘dislike’ both documents by the total amount of ‘likes’ and ‘dislikes’ of *A* or *B*. In the example depicted in Figure 3, two users like both documents, none dislikes them and the total amount equals five. The result (0.4) is the overlap that *A* and *B* share and may be interpreted as their similarity score. To leverage the content and the collaborative information, we combine both similarity values using a weighted average (80% content-based + 20% user-based) as proposed by Al-Hassan et al. (2015). The final score allows to generate recommendations, by suggesting the most similar documents for each currently viewed document to a user.

4 Evaluation and User Interface of the Artefact

The evaluation of our system is conducted twofold, as we aim to apply traditional recommendation system evaluation and approaches from KM research. First, we need quantifiable factors to determine the quality of the recommendations and thus of the proposed similarity calculation itself. Precision and Recall, for example, are well-established metrics to assess recommendation system quality by comparing the user history of ‘liked’ documents with provided recommendations (Cremonesi et al., 2010). Second, the user acceptance of a KM system plays a major role for the success of a KM initiative (Bals et al., 2007). IS research has developed well-established technology acceptance models (e.g. TAM, UTAUT) (Davis, 1989; Venkatesh et al., 2003) that measure factors influencing the acceptance of technology. Several studies have adapted these models for the evaluation of KM systems (Bals et al., 2007; Chen and Huang, 2010; Lin et al., 2004). To conduct our own user acceptance study, we are currently implementing a prototypical system at a Big 4 accounting company. To this end, we put forward a questionnaire aligned at the model of Lin et al. (2004). The authors include the perceived user-friendliness, individual factors (e.g. attitude), and organizational factors (e.g. the system’s fit into the everyday work context) to better measure the perceived usefulness (and thus acceptance) of KM systems. We furthermore intend to incorporate influencing factors (e.g. top-management support) drawn from Bals et al. (2007). As prerequisite for a user-centric evaluation, we provide a prototypical web-based user interface (see Figure 4), which offers (i) keyword search with filtering options, (ii) a detailed view of the retrieved documents containing the extracted descriptors, and (iii) a list of the most similar documents. Therefore, each individual user needs a user account, which allows to map his or her preferences to the database (cf. section 3.3) and to receive personalized proactive alerts in case new similar documents are crawled. Furthermore, we allow users to edit the automatically extracted keywords and references by clicking on the button “View entire list”.

Finally, after carrying out the system-centric and user-centric evaluation, we have to assess the benefit of the proposed legal KM system. However, this is a difficult task. Jennex et al. (2007) describe KM success as a multidimensional construct including process and outcome measures. The aim of our final evaluation is to assess both: (i) Does the system improve legal research processes, which are currently time-consuming, in terms of efficiency, and (ii) does the artefact improve our practice partner’s outcome in terms of quality and response speed (e.g. more comprehensive reports, enhanced recommendations for action related to legal changes)? We intent to employ a framework introduced by Fischer et al. (2011), who calculate the return on investment of a KM system using predefined metrics (e.g. time saved per employee) to estimate its value.



Search for documents by keyterms:

Results: 29 relevant matches found

Bewertung-Vorratsvermoegen-Lifo-Methode

DIREU2014-994658_En

Title: [...]entry and residence of third-country nationals in the framework of an intra- corporate transfer Original Language: English
Category: Regulation, Directive Publishing-Year 2014 Source EU

Extracted Entities and Keywords:
<Member States>$_{\text{[Organization]}}$ <EUROPEAN PARLIAMENT>$_{\text{[Organization]}}$ <social security systems>$_{\text{[FieldTerminology]}}$ <economic development>$_{\text{[Country]}}$ <public health>$_{\text{[FieldTerminology]}}$ Denmark>$_{\text{[Country]}}$ <insurance coverage>$_{\text{[FieldTerminology]}}$ <intra-corporate transferee>$_{\text{[Keyword]}}$ <trainee employees>$_{\text{[Keyword]}}$ <equal treatment>$_{\text{[Keyword]}}$ <national law>$_{\text{[Keyword]}}$ <fair competition>$_{\text{[Keyword]}}$

Extracted Legal References:
<Directive 96/71/EC> <Directive 2005/36/EC>li> <Regulation 562/2006/EC> <Directive 96/71/EC Article 1> <Regulation 883/2004/EC Article 3> <Regulation 1231/2010/EU> <Directive 2005/71/EC> <Regulation 1030/2002/EC Article 4>

Most similar documents:

-	Title	Shared Entities	Shared Ref	-
□	DIREU2014-32014_En	31	4	Details
□	[...]-and-Social-Conditions2006-99564	42	0	Details

(i) Search
Provides the possibility for keyword-based document search. In addition, the results can be filtered based on Source, Year, Language, and Country.

(ii) Details
Presents relevant matches and enables to expand a detail view for each document. This allows to assess the content of each document by the extracted entities, keywords, and references. (cf. Section 3.2) In addition, the button 'View entire List' provides functionality for manually editing the extracted features.

(iii) Recommendations
Shows the most similar documents to the currently selected document. The similarity is calculated according to section 3.3

Figure 4. User Interface of the Legal Knowledge Management System

5 Expected Contribution and Future Work

In this paper, we propose to leverage the potential of graph databases, text mining, and recommender systems for the retrieval and collaborative management of legal and legislative documents. We take the perspective that knowledge is a condition of access to information (McQueen, 1998) and show that our artefact fulfils the requirements for KM systems to create, store, transfer, and apply knowledge in an organization (Alavi and Leidner, 2001). The raw data collected by the web crawler or uploaded by users is transformed to information by categorization, automatic keyword annotation, and the possibility for user-based keyword editing. The storage and retrieval functionalities benefit from the advantages of a scalable graph database, which contains the relations between documents, features, and users. The transfer of implicit knowledge of the users is facilitated by the collaborative recommendation component, which generates document recommendations based on the relevance feedback of other users. Finally, the web-based user interface supports the application of knowledge at a consultancy firm. We expect that the artefact improves the processes of legal research, collaboration, and coordination in terms of quality and efficiency through timely and automatic routing of work-related documents. To find support for this assumption, we are currently working on a rigorous evaluation of the system which poses a difficult and time-consuming task. For reliable results, the trial period must involve a sufficient number of participants and consider their level of expertise, as well as various relevant theories (e.g. prospect theory). After completing the evaluation several opportunities for further research and functional improvements of the KM artefact arise. As the similarity calculation between legal documents is of fundamental importance, more sophisticated feature extraction methods hold the potential to improve the results of the entire system. Moreover, it would be conceivable that users with different information needs call for different document features. Therefore, the presented legal KM system can be extended by determining user-specific feature weightings for the similarity calculation (cf. Section 3.3). If, for example, a user's area of responsibility requires working more often with documents from a particular source, machine learning algorithms can be employed to take such characteristics into account.

References

Adomavicius, G. and Tuzhilin, A. (2005). “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions”, *IEEE transactions on knowledge and data engineering*, 17(6), pp. 734–749.

Agnoloni, T. and Tiscornia, D. (2010). “Extracting Normative Content from Legal Texts”, *Proceedings of the MCIS*, p. 4.

Alavi, M. and Leidner, D.E. (2001). “Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues”, *MIS quarterly*, 25(1), pp. 107–136.

Al-Hassan, M., Lu, H. and Lu, J. (2015). “A semantic enhanced hybrid recommendation approach: A case study of e-Government tourism service recommendation system”, *Decision Support Systems*, 72, pp. 97–109.

Angles, R. and Gutierrez, C. (2008). “Survey of graph database models”, *ACM Computing Surveys (CSUR)*, 40(1), p. 1.

Antonellis, I., Molina, H.G. and Chang, C.C. (2008). “Simrank++: query rewriting through link analysis of the click graph”, *Proceedings of the VLDB Endowment*, 1(1), pp. 408–421.

Antonini, A., Boella, G., Hulstijn, J. and Humphreys, L. (2013). “Requirements of legal knowledge management systems to aid normative reasoning in specialist domains”, *JSAI International Symposium on Artificial Intelligence.* Springer, pp. 167–182.

Balahur, A. and Turchi, M. (2014). “Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis”, *Computer Speech & Language*, 28(1), pp. 56–75.

Bals, C., Smolnik, S. and Riempp, G. (2007). “Assessing user acceptance of a knowledge management system in a global bank: Process analysis and concept development”, *40th Annual Hawaii International Conference on System Sciences (HICSS).* IEEE, 203c-203c.

Begoña, M.-F. and Carmen, P.-S. (2011). “Knowledge construction and knowledge sharing: a Wiki-based approach”, *Procedia-Social and Behavioral Sciences*, 28, pp. 622–627.

Bianchi, M., Draoli, M., Gambosi, G., Pazienza, M.T., Scarpato, N. and Stellato, A. (2009). “ICT tools for the discovery of semantic relations in legal documents”, *Proceedings of the 2nd International Conference on ICT Solutions for Justice (ICT4Justice).*

Boella, G., Di Caro, L., Rispoli, D. and Robaldo, L. (2013). “A system for classifying multi-label text into EuroVoc”, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law.* ACM, pp. 239–240.

Boella, G., Hulstijn, J., Humphreys, L., Janssen, M. and van der Torre, L. (2012). “Towards Legal Knowledge Management Systems for Regulatory Compliance”, *Proceedings of the IX Conference of the Italian Chapter of Association for Information Systems.* itAIS, Rome, Italy.

Boella, G., Humphreys, L., Martin, M., Rossi, P. and van der Torre, L. (2011). “Eunomos, a legal document and knowledge management system to build legal services”, *International Workshop on AI Approaches to the Complexity of Legal Systems.* Springer, pp. 131–146.

Carreras, X., Marquez, L. and Padró, L. (2002). “Named entity extraction using adaboost”, *Proceedings of the 6th conference on Natural language learning-Volume 20.* Association for Computational Linguistics, pp. 1–4.

Chen, H.-R. and Huang, H.-L. (2010). “User acceptance of mobile knowledge management learning system: Design and analysis”, *Educational Technology & Society*, 13(3), pp. 70–77.

Chen, X., Liu, Z. and Sun, M. (2014). “A Unified Model for Word Sense Representation and Disambiguation”, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* Doha, Qatar, pp. 1025–1035.

Cheng, H., Lu, Y.-C. and Sheu, C. (2009). “An ontology-based business intelligence application in a financial knowledge management system”, *Expert Systems with Applications*, 36(2), pp. 3614–3622.

Cremonesi, P., Koren, Y. and Turrin, R. (2010). "Performance of recommender algorithms on top-n recommendation tasks", *Proceedings of the fourth ACM conference on Recommender systems.* ACM, pp. 39–46.

Davis, F.D. (1989). "Perceived usefulness, perceived ease of use, and user acceptance of information technology", *MIS quarterly*, 13(3), pp. 319–340.

Deshpande, M. and Karypis, G. (2004). "Item-based top-n recommendation algorithms", *ACM Transactions on Information Systems (TOIS)*, 22(1), pp. 143–177.

Diligenti, M., Coetzee, F., Lawrence, S., Giles, C.L., Gori, M. and others (2000). "Focused Crawling Using Context Graphs", *Proceedings of the 26th VLDB Conference*, pp. 527–534.

Erdmann, M., Ikeda, K., Ishizaki, H., Hattori, G. and Takishima, Y. (2014). "Feature based sentiment analysis of tweets in multiple languages", *International Conference on Web Information Systems Engineering.* Springer, pp. 109–124.

Fischer, N., Hertlein, M., Smolnik, S. and Jennex, M.E. (2011). "Measuring Value of Knowledge-Based Initiatives-Evaluation of Existing Models and Development of a New Measurement Framework", *2011 44th Hawaii International Conference on System Sciences (HICSS).* IEEE, pp. 1–10.

Francesconi, E. and Passerini, A. (2007). "Automatic classification of provisions in legislative texts", *Artificial Intelligence and Law*, 15(1), pp. 1–17.

Gao, J., He, X., Yih, W.-t. and Deng, L. (2013). "Learning semantic representations for the phrase translation model", *arXiv preprint arXiv:1312.0482*.

Guo, J., Che, W., Wang, H. and Liu, T. (2014). "Learning Sense-specific Word Embeddings By Exploiting Bilingual Resources", *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pp. 497–507.

Hagen, L., Uzuner, Ö., Kotfila, C., Harrison, T.M. and Lamanna, D. (2015). "Understanding Citizens' Direct Policy Suggestions to the Federal Government: A Natural Language Processing and Topic Modeling Approach", *48th Hawaii International Conference on System Sciences (HICSS).* IEEE, pp. 2134–2143.

He, W. and Yang, L. (2016). "Using wikis in team collaboration: A media capability perspective", *Information & Management*, 53(7), pp. 846–856.

Hevner, A.R., March, S.T., Park, J. and Ram, S. (2004). "Design science in information systems research", *MIS quarterly*, 28(1), pp. 75–105.

Hsu, J. (2016). "For Sale: Deep Learning". Available at: ieeexplore.ieee.org/iel7/6/7524147/07524158.pdf.

IJntema, W., Goossen, F., Frasincar, F. and Hogenboom, F. (2010). "Ontology-based news recommendation", *Proceedings of the 2010 EDBT/ICDT Workshops.* ACM, p. 16.

Jackson, P., Al-Kofahi, K., Tyrrell, A. and Vachher, A. (2003). "Information extraction from case law and retrieval of prior cases", *Artificial Intelligence*, 150(1), pp. 239–290.

Jeh, G. and Widom, J. (2002). "SimRank: a measure of structural-context similarity", *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, pp. 538–543.

Jennex, M.E., Smolnik, S. and Croasdell, D. (2007). "Towards defining knowledge management success", *40th Annual Hawaii International Conference on System Sciences (HICSS).* IEEE, 193c-193c.

Jetzek, T., Avital, M. and Bjørn-Andersen, N. (2013). "Generating value from open government data", *Proceedings of the 34th International Conference on Information Systems*, pp. 1–20.

Jiang, R., Banchs, R.E. and Li, H. (2016). "Evaluating and Combining Named Entity Recognition Systems", *Proceedings of the Sixth Named Entities Workshop (NEWS)*, p. 21.

Kankanhalli, A., Lee, O.-K.D. and Lim, K.H. (2011). "Knowledge reuse through electronic repositories: A study in the context of customer service support", *Information & Management*, 48(2), pp. 106–113.

Kankanhalli, A., Tan, B.C.Y. and Wei, K.-K. (2005). "Contributing knowledge to electronic knowledge repositories: an empirical investigation", *MIS quarterly*, 29(1), pp. 113–143.

Kiniti, S. and Standing, C. (2013). “Wikis as knowledge management systems: issues and challenges”, *Journal of Systems and Information Technology*, 15(2), pp. 189–201.

Knackstedt, R., Eggert, M., Heddier, M., Chasin, F. and Becker, J. (2013). “The Relationship Of Is And Law-The Perspective Of And Implications For IS Research”, *Proceedings of the 21st European Conference on Information Systems (ECIS)*, pp. 1–12.

Knackstedt, R., Heddier, M. and Becker, J. (2014). “Conceptual modeling in law: An interdisciplinary research agenda”, *Communications of the Association for Information Systems*, 34(1), p. 36.

Ko, D.-G., Kirsch, L.J. and King, W.R. (2005). “Antecedents of knowledge transfer from consultants to clients in enterprise system implementations”, *MIS quarterly*, 29(1), pp. 59–85.

Lagos, N., Segond, F., Castellani, S. and O’Neill, J. (2010). “Event extraction for legal case building and reasoning”, *International Conference on Intelligent Information Processing.* Springer, pp. 92–101.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C. (2016). “Neural architectures for named entity recognition”, *arXiv preprint arXiv:1603.01360*.

Lauly, S., Larochelle, H., Khapra, M., Ravindran, B., Raykar, V.C. and Saha, A. (2014). “An autoencoder approach to learning bilingual word representations”, *Advances in Neural Information Processing Systems*, pp. 1853–1861.

Li, M., Liu, L. and Li, C.-B. (2011). “An approach to expert recommendation based on fuzzy linguistic method and fuzzy text classification in knowledge management systems”, *Expert Systems with Applications*, 38(7), pp. 8586–8596.

Li, X., Montazemi, A.R. and Yuan, Y. (2006). “Agent-based buddy-finding methodology for knowledge sharing”, *Information & Management*, 43(3), pp. 283–296.

Lin, C., Hu, P.J.-H. and Chen, H. (2004) “Technology implementation management in law enforcement: COPLINK system usability and user acceptance evaluations”, *Social Science Computer Review*, 22(1), pp. 24–36.

Lin, F.-R., Chou, S.-Y., Liao, D. and Hao, D. (2015). “Automatic content analysis of legislative documents by text mining techniques”, *48th Hawaii International Conference on System Sciences (HICSS).* IEEE, pp. 2199–2208.

Lin, F.-R., Huang, Y.-t. and Liao, D. (2012). “Incrementally Clustering Legislative Interpellation Documents”, *45th Hawaii International Conference on System Science (HICSS).* IEEE, pp. 2521–2530.

Lippi, M. and Torroni, P. (2016). “Argumentation mining: State of the art and emerging trends”, *ACM Transactions on Internet Technology (TOIT)*, 16(2), p. 10.

Maat, E. de, Krabben, K. and Winkels, R. (2010). “Machine Learning versus Knowledge Based Classification of Legal Texts”, *JURIX*, pp. 87–96.

McQueen, R. (1998). “Four views of knowledge and knowledge management”, *Proceedings of the Fourth Americas Conference on Information Systems (AMCIS)*, p. 204.

Mourtzis, D. and Doukas, M. (2014). “Knowledge capturing and reuse to support manufacturing of customised products: A case study from the mould making industry”, *Procedia CIRP*, 21, pp. 123–128.

Nadeau, D. and Sekine, S. (2007). “A survey of named entity recognition and classification”, *Lingvisticae Investigationes*, 30(1), pp. 3–26.

Nastase, V. and Strube, M. (2013). “Transforming Wikipedia into a large scale multilingual concept network”, *Artificial Intelligence*, 194, pp. 62–85.

Navigli, R. and Ponzetto, S.P. (2012). “BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network”, *Artificial Intelligence*, 193, pp. 217–250.

Ni, X., Sun, J.-T., Hu, J. and Chen, Z. (2011). “Cross lingual text classification by mining multilingual topics from wikipedia”, *Proceedings of the fourth ACM international conference on Web search and data mining.* ACM, pp. 375–384.

Palmirani, M., Brighi, R. and Massini, M. (2003). “Automated extraction of normative references in legal texts”, *Proceedings of the 9th international conference on Artificial intelligence and law.* ACM, pp. 105–106.

Pant, G., Srinivasan, P. and Menczer, F. (2004). “Crawling the web”, in *Web Dynamics:* Springer, pp. 153–177.

Pecina, P., Dušek, O., Goeuriot, L., Hajič, J., Hlaváčová, J., Jones, G.J.F., Kelly, L., Leveling, J., Mareček, D., Novák, M. and others (2014). “Adaptation of machine translation for multilingual information retrieval in the medical domain”, *Artificial intelligence in medicine*, 61(3), pp. 165–185.

Rizzo, G. and Troncy, R. (2011). “Nerd: evaluating named entity recognition tools in the web of data”, *Workshop on Web Scale Knowledge Extraction*, pp. 1–16.

Rodrigo, Á., Pérez-Iglesias, J., Peñas, A., Garrido, G. and Araujo, L. (2013). “Answering questions about European legislation”, *Expert Systems with Applications*, 40(15), pp. 5811–5816.

Rodriguez-Enriquez, C.A., Alor-Hernandez, G., Mejia-Miranda, J., Sanchez-Cervantes, J.L., Rodriguez-Mazahua, L. and Sanchez-Ramirez, C. (2016). “Supply chain knowledge management supported by a simple knowledge organization system”, *Electronic Commerce Research and Applications*, 19, pp. 1–18.

Santos, C.N.d. and Guimaraes, V. (2015). “Boosting named entity recognition with neural character embeddings”, *Proceedings of the Fifth Named Entity Workshop (NEWS).*

Savvas, I. and Bassiliades, N. (2009). “A process-oriented ontology-based knowledge management system for facilitating operational procedures in public administration”, *Expert Systems with Applications*, 36(3), pp. 4467–4478.

Schaffert, S. (2006). “IkeWiki: A semantic wiki for collaborative knowledge management”, *15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE’06).* IEEE, pp. 388–396.

Tello-Leal, E., Rios-Alvarado, A.B. and Diaz-Manriquez, A. (2015). “A Semantic Knowledge Management System for Government Repositories”, *26th International Workshop on Database and Expert Systems Applications (DEXA).* IEEE, pp. 168–172.

Tran, O.T., Le Nguyen, M. and Shimazu, A. (2013). “Reference resolution in legal texts”, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law.* ACM, pp. 101–110.

Trappey, A.J.C. and Trappey, C.V. (2008). “An R&D knowledge management method for patent document summarization”, *Industrial Management & Data Systems*, 108(2), pp. 245–257.

Tsoi, A.C., Frosali, D., Gori, M., Hagenbuchner, M. and Scarselli, F. (2003). “A Simple Focused Crawler”, *World Wide Web Conference,* Budapest.

Varga, A. and Edmonds, A.N. (2016). “Multilingual extraction and editing of concept strings for the legal domain”, *Advances in Computer Science: an International Journal*, 5(4), pp. 18–23.

Venkatesh, V., Morris, M.G., Davis, G.B. and Davis, F.D. (2003). “User acceptance of information technology: Toward a unified view”, *MIS quarterly,* 27(3), pp. 425–478.

Wagner, C. (2004). “Wiki: A technology for conversational knowledge management and group collaboration”, *The Communications of the Association for Information Systems*, 13(1), p. 58.

Wei, C.-P., Yang, C.C. and Lin, C.-M. (2008). “A Latent Semantic Indexing-based approach to multilingual document clustering”, *Decision Support Systems*, 45(3), pp. 606–620.

Wu, H., Gordon, M.D. and Fan, W. (2010). “Collective taxonomizing: A collaborative approach to organizing document repositories”, *Decision Support Systems*, 50(1), pp. 292–303.

Zhen, L., Huang, G.Q. and Jiang, Z. (2010). “An inner-enterprise knowledge recommender system”, *Expert Systems with Applications,* 37(2), pp. 1703–1712.