Matching Knowledge Users with Knowledge Creators using Text Mining Techniques

Abdulrahman Al-Haimi
Department of Management Sciences, University of Waterloo, Waterloo, Ontario, Canada

Keywords: Text Mining, Matching, Knowledge Users, Knowledge Creators, Clustering, Classification, Management, Experimentation, Exploration, Performance.

Abstract: Matching knowledge users with knowledge creators from multiple data sources that share very little similarity in content and data structure is a key problem. Solving this problem is expected to noticeably improve research commercialization rate. In this paper, we discuss and evaluate the effectiveness of a comprehensive methodology that automates classic text mining techniques to match knowledge users with knowledge creators. We also present a prototype application that is considered one of the first attempts to match knowledge users with knowledge creators by analyzing records from Linkedin.com and BASE-search.net. The matching procedure is performed using supervised and unsupervised models. Surprisingly, experimental results show that K-NN classifier shows a slight performance improvement compared to its competition when evaluated in a similar context. After identifying the best-suited methodology, system architecture is designed. One of the main contributions of this research is the introduction and analysis of a novel prototype application that attempts to bridge the gap between research performed in industry and academia.

1 INTRODUCTION

Research commercialization has been the focus of leading universities and public research institutes worldwide in an attempt to accelerate innovation (Dooris 1989; Etzkowitz and Peters 1991). A variety of programs have been created (e.g., technology transfer offices, incubator and accelerator centres) to primarily boost the rate of commercialized research by matching knowledge creators with knowledge users. While helpful, these programs suffer from lengthy and inefficient processes in general (Siegel et al. 2004; Swamidass and Vulasa 2009). Recent studies and reports show that countries that invest billions of dollars on research still do not achieve high rates of commercialized research (Swamidass and Vulasa 2009; Council of Canadian Academies 2012). In other words, majority of existing research commercialization programs seem not as effective as expected.

We believe that the commercialization rate can improve noticeably if there exists a comprehensive online application that discovers hidden connections between information created by both knowledge users and knowledge creators. Shedding light on those hidden connections helps knowledge creators (e.g., researchers) to easily measure the level of market demand for their research, measure the level of global research efforts that tackle similar problems, and locate potential commercialization opportunities. It also helps knowledge users (e.g., business organizations) to locate sources of expert knowledge that can help to develop a portfolio of innovative products and services. Ultimately, such an online tool helps in answering questions like who the market players are for certain research topics, how big the market is in terms of supply and demand, who the top matches are, and what the most sustainable match is for a given knowledge creator or knowledge user. Knowing answers to questions like these helps in discovering and aligning research interests of knowledge creators and knowledge users.

Matching knowledge creators with knowledge users is difficult due to the lack of a commonly defined procedure for this matching problem and the lack of simplified operational measures (Bozeman, 2000; Etzkowitz, 2002). The existing matching procedures vary significantly depending on whether they are initiated from academia, industry, or

Al-Haimi A..
Matching Knowledge Users with Knowledge Creators using Text Mining Techniques.
DOI: 10.5220/0004942000050014
In *Proceedings of 3rd International Conference on Data Management Technologies and Applications* (DATA-2014), pages 5-14
ISBN: 978-989-758-035-2

government (Bozeman, 2000). Given that classic text mining techniques are capable of solving the matching problem (Kannan et al., 2011; Liu et al., 2011), we adopt some of them to develop a standard methodology that has no bias to academia, industry, or government. Applying classic techniques enables establishing a baseline for future research since this paper represents one of the early attempts to develop an application that matches knowledge creators with knowledge users.

A major advantage about the methodology proposed here over existing commercialization methods is that a match can be realized without the need for offline information search. Analyzing and mining information available in digital academic databases and job posting databases can reveal the amount of information necessary to generate effective matches. Thanks to technological discoveries in the field of text mining, text documents can now be automatically analyzed in order to find a list of a top-k keywords that best describes any given text document based on its content. Today, text documents can automatically be classified with minimal human-intervention.

While online data makes data retrieval an easy process, online data almost always scores poorly in terms of data quality, especially when retrieving data from multiple sources. For example, creating a single, high quality, dataset that contains online data from various sources about a single product can be difficult due to the potentially large amount of data transformation needed. Creating one single, high quality, dataset about multiple products becomes a challenge.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 discusses the proposed methodology. Section 4 presents the prototype application. Section 5 describes evaluates the performance of the proposed methodology and application. Finally, conclusions and directions for future work are discussed in section 6.

2 RELATED WOK

Research commercialization is operationalized and measured by the rate of (a) patenting and licensing, (b) new venture and spin-off creation, and (c) sponsored research collaborations (Rogers et al., 2001; Nordfors et al., 2003). However, these metrics provide limited insights to someone who needs to analyze the performance of a given research commercialization process from start to end, as it focuses only on end results. Prior research lacks a comprehensive analysis about the most reliable measures that can be used to match knowledge users with knowledge creators (Bozeman, 2000; Etzkowitz, 2002). However, they indirectly identify a number of measures that can be used such as publications, patents, licenses, publication or patent citations, and job postings (Nordfors et al., 2003; Karlsson, 2004; Campbel and Swigart, 2014). In this research, we attempt to obtain the matches by finding similarities between publications and job postings.

Though the matching problem has been well studied (Bilenko and Mooney 2003; Elmagarmid et al., 2007; Dorneles et al., 2010; Kannan et al., 2011), two general assumptions from past research do not hold in the matching problem discussed here. First, past research explores the matching problem where records are syntactically different, but refer to the same object. Finding matching records from two different academic databases describing one unique research paper can be taken as an example. Second, properties of these records can either be well structured or non-structured (available in text format), but not both.

Previous studies on record linkage and duplicate detection assume records to be properly structured and well segmented (Newcombe et al., 1959; Fellegi and Sunter, 1969). On the other hand, studies on natural language processing pay more attention to finding similar records that are available in unstructured text (Mitkov, 2002). Although, the problem identified in this research is different to some extent than the problems discussed in past research, they are related and should be taken in consideration. For example, research publications usually contain both structured and unstructured information. Structured information can be found in attributes like title, keywords, authors, and publishing agency. In contrast, unstructured information can be located in attributes like research abstract.

Text mining studies that focus on matching different populations or datasets use a number of techniques such as vector-based similarity measures, clustering, classification or a variation of these techniques (Chung, 2004; Zhou et al., 2007). Other studies approach this matching problem by examining techniques like ontologies and semantics (Maedche and Staab, 2001; Antezana et al., 2009). Matching records that are syntactically different, but refer to the same object has been referred to as entity resolution, identity uncertainty, object identification, duplicate detection, and record linkage (Bilenko and Mooney, 2003; Elmagarmid et al., 2007; Dorneles et

al., 2010; Kannan et al., 2011). Record matching can generally be done using structure-based techniques or content-based techniques (Dorneles et al., 2010). Research in this field has focused primarily on three aspects: (a) selecting possible matching records by blocking impossible matches (e.g., if no common tokens can be found), (b) calculating similarity measures between possible matching records, and lastly (c) filtering out records that do not meet a pre-defined similarity score threshold t.

Implementation of a matching system has been carried out through supervised and unsupervised machine learning models. While a number of studies find supervised machine learning models perform better than their counterparts (Yang and Liu, 1999; Li and Yang, 2003; Özgür et al., 2005), they are very costly to maintain since they need a balanced and updated training dataset to keep results effective (Köpcke et al., 2010). Others believe that hybrid approaches can provide superior performance (Xiang et al., 2012). Bilenko and Mooney (2003) and Köpcke et al. (2010) evaluated the performance levels of supervised and unsupervised classification models using real-world cases. Based their findings, we believe that a hybrid approach that relies on an unsupervised clustering model and a supervised classification model to automatically cluster and match different records is expected to provide effective and efficient results.

3 METHODOLOGY

This section discusses a text mining procedure aims to retrieve records from two different data sources to measure market supply and demand for research and innovation to ultimately match knowledge users with knowledge creators. The following subsections discuss this procedure in detail.

3.1 Retrieval

Information about knowledge creators is retrieved from a well-organized open-access academic database called BASE-search.net, which crawls thousands of digital libraries from authentic sources to search for research publications. BASE claims to have access to over 60 million publications from more than 3000 sources (Bielefeld University, 2014). Information about knowledge users is collected from LinkedIn job-posting database, with more than 350,000 job posts worldwide. With this information in mind, these two sites are going to be the data sources we use to illustrate the effectiveness of the proposed methodology to produce the matches.

The study selected these two sources primarily for two reasons: (a) data representation, and (b) unrestricted data access privileges. Information stored in these data sources contains metadata that can be extracted to improve the matching performance. In addition, information is stored in a format that allows easy metadata extraction using XPath or Regex. More importantly, these sources are openly accessible. These features make the sources more attractive for the purpose of evaluating the performance of the proposed method and prototype.

Based on a user-defined search query, a web crawler crawls web pages that match the crawling criteria to follow and store only pages that represent job posts or research publications within the defined data sources using regular expression (Regex). The top k relevant results from both sources are then passed over to the pre-processing stage, where text analysis of each web page is performed in order to find the most important and least important features (i.e., words) that describe those pages.

3.2 Pre-Processing

Two important tasks are performed here: (a) metadata extraction, and (b) feature extraction. The first task parses a standard HTML page using Regex and XPath to store any structured metadata that can be found (e.g., title, authors, organization,) to allow for further analysis.

Feature extraction is more complex than the first task. It consists of a number of subtasks. First, text extraction, which entails looking at a selected part of a web page based on its content and removes any HTML tags that can be found. This results in a group of sentences that best describes the content of a given web page. Second, tokenization, which separates sentences into a list of words or parts of words. Third, filtering, which removes tokens by length and by function. Following suggestions from past research, words that are less than 2-character long or more than 25-character long are filtered out (Ertek et al., 2013; Ramesh, 2014). Also, stop words (e.g., about, from) are filtered out. Fourth, remaining tokens are transformed to lower-case in order to aggregate similar words that are written in different forms. Lastly, stemming of tokens is performed to increase similarity between different records (web pages) by reducing extracted tokens to their simplest form removing suffixes that might be attached to any token. This also allows for feature vector size reduction. This study incorporates a stemming

algorithm for English developed by Porter (1980), which is widely adopted (Ramesh, 2014).

Filtering is also performed by pruning features that have either low frequency (less than 20%) or high frequency (more than 80%) of occurrence in the record collection. For example, if the extracted feature "machine" occurs in 5 web pages out of 100 crawled pages, it will be filtered out. This technique is adopted in past research to remove features from the corpus that do not help to identify certain records in the collection (Deerwester et al., 1990; Carmel et al., 2001).

The final step in pre-processing is to create a feature vector set combining the final list of features. Records in the collection are distinguished based on their corresponding $\mathrm{tf} \cdot \mathrm{idf}$ weights. Borrowed from the field of information retrieval, the $\mathrm{tf} \cdot \mathrm{idf}$ weighting scheme (see equation 1) is used here to assign weights to features based on textual contents of records. Cohen (1998) separates each string σ into words and each word w is assigned a weight where the term frequency for word w in a single record is $\mathrm{tf_w} = \frac{\mathrm{f_w}}{|\mathrm{W}|}$ and the inverse document frequency $\mathrm{idf_w} = \frac{|\mathrm{R}|}{\mathrm{n_w}}$. The frequency of word w in this record is $\mathrm{f_w}$, W is the total count of tokens in the same record, R is the total number of records in the collection, and $\mathrm{n_w}$ is the number of records that contain the word w.

$$\mathcal{A}_\sigma(w) = \log(tf_w + 1) \cdot \log(idf_w) \quad (1)$$

The $\mathrm{tf} \cdot \mathrm{idf}$ weight for word w in a record is considered high (maximum of 1.0) if w appears a large number of times in that record, while being a slightly rare term in the collection of records. For example, for a data collection that contains 1000 records about university names, relatively infrequent terms such as 'Waterloo' or 'Guelph' have higher $\mathrm{tf} \cdot \mathrm{idf}$ weights than more frequent terms such as 'University'.

3.3 Transformation

The process of extracting metadata, which we introduced earlier on, aims to enable the analysis of structured data to ultimately incorporate these data to search for the best matches. However, due to the noise that usually exists in metadata, this cannot be done until data records are cleaned. To illustrate, imagine that the collection of records available to you contains metadata about the language of these records. It is possible to find that records do not follow a unified structure. English language can be referred to as 'en', 'ENG', 'E', 'ENGLISH', etc. To avoid these issues, extracted metadata is transformed in a way to create a unified structure that allows for better performance.

3.4 Joining

To be able to match knowledge users with knowledge creators, feature vector sets that represent both knowledge users and knowledge creators are joined in one superset. Cosine similarity (see equation 2) combined with the $\mathrm{tf} \cdot \mathrm{idf}$ weighting scheme are used here to compute the similarity between two records of the same class or different class (e.g., publication vs. publication, publication vs. job post). To account for relative document word count, (a) only relevant publication information (e.g., title, authors, abstract, keywords) available in Base-search.net domain is analyzed as apposed to analyzing the entire publication document, (b) the $\mathrm{tf} \cdot \mathrm{idf}$ weighting scheme normalizes term frequencies in each document based on how important a word is to that document. Words that are least important to a document are given a value very close to zero. These words are then dropped out of the list. The cosine similarity of σ_1 and σ_2 is defined as follows:

$$\mathcal{S}im(\sigma_1, \sigma_2) = \frac{\sum_{j=1}^{|R|} \mathcal{A}_\sigma(j) \cdot \mathcal{A}_\sigma(j)}{\|\mathcal{A}_{\sigma 1}\|_2 \cdot \|\mathcal{A}_{\sigma 2}\|_2} \quad (2)$$

Cosine similarity is very effective for when used with a large corpus of words and it is insensitive to the position of words, thus allowing natural word moves and swaps (e.g., 'Don Kawasaki' is equivalent to 'Kawasaki, Don'). Also, frequent words virtually do not affect the similarity of the two strings due to the low $\mathrm{tf} \cdot \mathrm{idf}$ weight of frequent words. For example, 'John Smith' and 'Mr. John Smith' would have similarity close to 1.0.

Similar records are placed in one cluster based on cosine similarity scores. The number of clusters is determined automatically using unsupervised clustering model X-Means clustering algorithm. Past research found that X-Means performed better and faster than the commonly used clustering algorithm, K-Means (Pelleg and Moore, 2000). One important difference to be pointed out here between the two is that K-Means requires users to identify the number of clusters in advance. X-Means, on the other hand, employs a different algorithm that learns from the feature vector set and creates the optimum number of clusters that can best differentiate records in the collection.

Since unsupervised clustering models usually lacks accuracy (Winkler, 2002), we adopt a two-level clustering design to initiate the matches. X-Means algorithm is used first to cluster and match records in the collection, then human experts are double check the results and make the required changes to maintain high quality results. With this in mind, experts can suggest similar records that should be clustered together based on their experience. Details from the two-level clustering are then pushed over to a supervised classification model to make the final matching of records. The next subsection explains the clustering procedure in more detail.

3.5 Evaluation and Matching

To ensure better cluster classification, Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) supervised models are evaluated first. The best performer is incorporated to cluster individual records in the collection. The process for clustering starts once records are stored in a database table when crawling web pages is completed. To ensure fast and accurate results, clustering records happens in two ways: (a) automatic clustering based on an unsupervised clustering model named X-Means, and (b) optional expert-generated clustering based on a supervised clustering model (e.g., SVM, K-NN). By combining unsupervised models with supervised ones to determine best matches, we expect to outperform traditional models (Li et al., 2009). In cases where there are no expert-generated models, clustering methods generated by X-Means algorithm is adopted. If there are expert-generated models, these models are taken in consideration (with higher weighting scheme) to classify records next time the process is initiated. Suggestions from human experts are recorded as a user feedback on the X-Means clustering method, which ultimately helps to improve the accuracy of X-Means predictive abilities.

4 SOLUTION

Based on the adopted methodology, this section showcases a prototype application as a first attempt to match knowledge users with knowledge creators using keyword search. The development of this application follows a system architecture that can be found in Figure 1. The following subsections discuss three modules of the prototype in mode detail.

4.1 Data Access

The prototype application accesses two sources of information. First, BASE-search.net, which contains academic publications that are assumed to measure, partly, the domain knowledge of knowledge creators. Second, LinkedIn job search engine, which contains job posts that are also assumed to measure, partly, knowledge needs of knowledge users. Since BASE-search.net hosts a number of academic publication types, this research explores only a selection of papers, articles, theses, reviews, and software records.

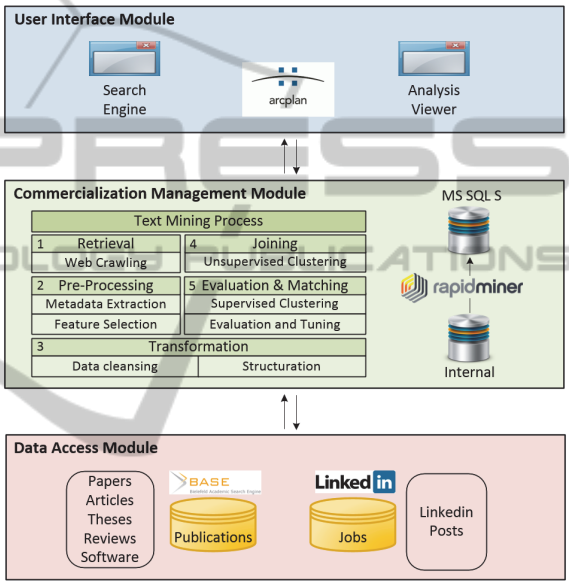

Figure 1: System Architecture.

4.2 Commercialization Management

This module is initiated once a user submits a search query about a subject of interest. Following the same methodology discussed in section 3, the module commences by retrieving related data using web crawling techniques. It is important to note that web crawlers are designed to search for the top 50 most relevant records from each source. The restriction is done for two reasons: (a) users normally do not read records beyond this number (Chitika Insights, 2013), and (b) it helps the computer that runs these processes perform fairly well.

One of the most important procedures that enable matching knowledge creators with knowledge users is evaluation and matching. To perform supervised clustering, a total of 100 records are retrieved and stored in an internal database. A training dataset, which is manually labelled by an expert, consists of

20 records from the dataset being classified. After clustering records, a new dataset is generated and stored in Microsoft SQL Server 2008 database table to be accessed later for data analysis. It is important to note that text mining processes that are discussed in this research have been implemented using a leading open-source data mining tool named RapidMiner (Mierswa et al., 2006). Figure 2 shows an overview of the commercialization management process that we have developed using RapidMiner.

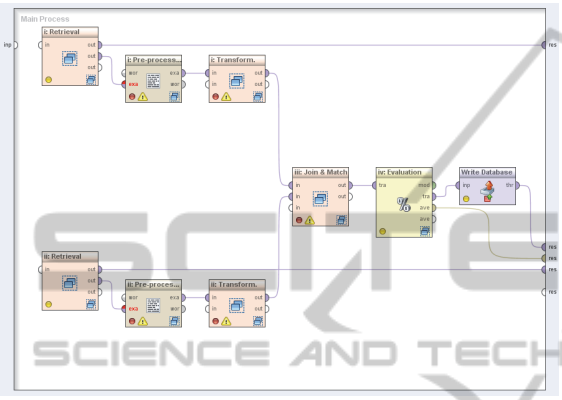

Figure 2: Text Mining Process Overview.

4.3 User Interface

This module describes the Graphical User Interface (GUI) designed to allow users of this application to interact with the other components of the system in a friendly and easy-to-use platform. The prototype application is named 'Innovation Base', developed using arcplan application designer tool. There are mainly two screens in Innovation Base. The first screen allows users to insert a keyword search, capable of reading Boolean arguments. This screen passes the search query to the text mining process defined in RapidMiner. The process starts by crawling the two data sources identified above, and pulling in records accordingly. The other screen, 'Analysis Viewer', depicts the output of the commercialization management module, particularly the result of classification, in a user-friendly format.

Analysis Viewer (see Figure 3) displays three types of information. First, information about the top ten business organizations and academic content providers which are most represented by the search keywords. Information displayed is sorted based on the number of aggregated records. Second, trend analysis can be done by evaluating information displayed in a timeline graph, which shows aggregated records based on their publishing date for knowledge users and knowledge creators.

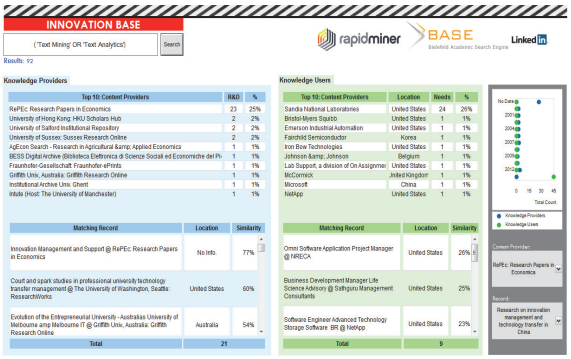

Figure 3: Analysis Viewer.

Finally, detailed information about the computed classification is presented in the lower half of this screen. This part displays three levels of detail in a single field: (a) title of a matching record, (b) name of academic content provider (e.g., HighWire Press) or business organization (e.g., Microsoft), and (c) name of author(s). Information about the place of origin of the records and calculated similarity measures are also displayed in the lower half of the screen.

5 EXPERIMENT

To test the validity of the identified methodology, we performed a comprehensive test using multiple search queries. To illustrate, one of the search queries ('Text Mining' OR 'Text Analytics') retrieved more than 7600 research publications from BASE-search.net database, and more than 200 job posts from LinkedIn job posting database. Top 50 relevant search results (from each of the sources) are chosen for matching.

Results of this experiment are evaluated using four metrics commonly used to evaluate and contrast performance of classifiers. First, Precision (P), which calculates the percentage of relevant records that are predicted by a classifier out of all predicted records. Second, Recall (R), which calculates the percentage of relevant records that are predicted by a classifier out of all relevant records. Third, F1-score, which uses P and R scores to calculate the accuracy of a classifier. Finally, execution time, which is used to measure the time a classifier takes to complete the classification task. It is important to note that of all the settings that are tested, only those that provided the best results are reported here.

Table 1: Supervised ML Model Performance (1st run).

Model	Class	P	R	F1	Exec. Time
K-NN (K=3)	Cluster_0	0.830	0.916	0.870	2 Sec.
	Cluster_1	0.913	0.823	0.856	
SVM	Cluster_0	0.977	0.895	0.934	2 Sec.
	Cluster_1	0.909	0.980	0.943	

Supervised machine learning models performed close to optimum after a single run of optimization. Table 1 shows the performance output of the first run before optimization. To improve performance, the paper includes more stemming rules and optimized feature selection. Table 2 represents the final performance output. Further testing and evaluations are based on results generated by the optimized (Table 2) model.

Table 2: Supervised ML Model Performance (2nd run).

Model	Class	P	R	F1	Exec. Time
K-NN (K=3)	Cluster_0	0.978	0.978	0.978	1 Sec.
	Cluster_1	0.980	0.980	0.980	
SVM	Cluster_0	0.940	1.00	0.969	1 Sec.
	Cluster_1	1.00	0.942	0.970	

A quick look at the F1-score of the tested models reveals that K-NN model outperforms SVM model by almost 1 per cent. This clearly shows the competitiveness of both models in terms of classifying text-based records that represent knowledge users and knowledge creators.

To visually illustrate the accuracy of these models, the Confusion Matrix measure is analyzed. This technique simulates the performance results of a supervised machine-learning model based on its ability to predict the right class for unclassified records.

Table 3: Confusion Matrix Measure (K-NN).

K-NN (K=3)	True Cluster_0	True Cluster_1
Pred. Cluster_0	46	1
Pred. Cluster_1	1	51

Analyzing other performance results shown in Table 3 and 4 provides additional useful details. For example, while having an accuracy of 98 per cent, K-NN model seems to have a difficulty in learning how to effectively classify records in any given class. On the other hand, while scoring an accuracy of 97 per cent, SVM model seems to effectively learn the classification rules associated with Cluster_0. It correctly identifies all records that originally belong to this cluster. However, it falls short in learning the appropriate classification rules to identify 3 records that belong to the other class. While both have positives and negatives, analyzing the confusion matrixes of SVM and K-NN do not provide inclusive evidence to determine which model provides better classification with this dataset.

Table 4: Confusion Matrix Metrics (SVM).

SVM	True Cluster_0	True Cluster_1
Pred. Cluster_0	47	3
Pred. Cluster_1	0	49

It is also worth mentioning that in the original experiment setup, a Part-Of-Speech (POS) tagger was included to filter out features that are not either nouns or adjectives. Due to performance issues we have decided to remove this tagger from the process. Table 5 shows the performance of the text mining process with and without a POS tagger.

Table 5: POS Tagger Performance Results.

	Number of Features	Exec. Time
POS Tagger	237	2:30 Min.
No POS Tagger	271	10 Sec.

Overall, the text mining process from crawling the two different data sources to classifying records into different clusters takes about 1:51 minutes in live mode where data is not stored locally in a hard disk. In offline mode, this process takes about 10 seconds from start to finish.

In the following subsection, a benchmarking performance measure is calculated to evaluate results generated here, and evaluate the goodness of the adopted methodology.

5.1 Benchmarking

In this subsection, we evaluate the validity of the adopted models using Silhouette validity index. This technique computes the silhouette width for each data record, average silhouette width for each cluster and overall average silhouette width for the total dataset (Rousseeuw, 1987). To compute the silhouettes width of data record i, the following formula is used:

$$S_i = \frac{b_i - a_i}{\max(a_i,\ b_i)} \quad (3)$$

where a_i is average dissimilarity of record i to all other records in the same cluster; b_i is minimum of average dissimilarity of record i to all records in other cluster. Dissimilarity between records is calculated using Cosine Similarity measure. A value of S_i close to 1.0 indicates that a record is assigned to the right cluster. If S_i is close to zero, it means that that record could be assigned to another cluster because it is equidistant from a number of clusters. If S_i is close to -1.0, it means that record is misclassified and lies somewhere in between the clusters. The overall average silhouette width for the entire dataset is the average S_i for all records in the dataset. The largest overall average silhouette indicates the best clustering and classification model (Rousseeuw, 1987).

Table 6: SVI Performance Results.

Model	Class	SVI	SVI (Avg)
K-NN (K=3)	Cluster_0	0.862	0.894
	Cluster_1	0.927	
SVM	Cluster_0	0.786	0.805
	Cluster_1	0.825	

Performance results in Table 6 shows that K-NN classifier is more accurate than SVM classifier for the given example. This result is somewhat surprising since it does not align with results from similar text mining research studies in the field of information retrieval (Yang and Liu, 1999; Li and Yang, 2003; Özgür et al., 2005). This result could be attributed to the limited sample size (n=100). SVM is proven to demonstrate very good results with a large sample size (Nidhi and Gupta, 2011). However, some research results suggest that SVM is not always the best classifier that can be used for text classification problems (Colas and Brazdil, 2006). In fact, K-NN is believed to be able to generate similar results under the 'right' implementation (Colas and Brazdil, 2006).

Generally, the experimental results reveal interesting insights. There is a high performance competition between K-NN and SVM in terms of the example dataset reported in this subsection. When models are evaluated under F1-score, K-NN presented better recall score on all cases by about one per cent. A similar conclusion can also be drawn by analyzing data from the confusion matrix. The noticeable feature about SVM classifier is the superior performance in predicting the correct classes for majority of records without the need for defining a constant number of centroids (k) before hand. This is important in cases where sample size is not limited to a specific number. Since the proposed solution is designed to allow users to identify the maximum number of records retrieved in online mode, this feature is important to be taken in consideration. At this point, evaluation results of SVI confirm that K-NN outperforms SVM. This means that K-NN is better in classifying records that maintain a high level of similarity. Measuring execution time reveals that both of these models complete the classification task at the same time.

Colas and Brazdil (2006) and Sokolova et al (2006) identify other measures that can also be used to evaluate the performance of classifiers, however, the measures we have reported here evaluate the performance of SVM and K-NN in many ways. At this point, it can safely be concluded that the above evaluated performance measures provide sufficient proof about the validity of the adopted methodology. Perhaps, new measures can be adopted in future work for comparative purposes. Based on these results, K-NN is considered the most effective and efficient classifier in the discussed context.

6 CONCLUSION

In this paper, we tackle the problem of matching knowledge users with knowledge creators using information extracted from digital academic databases and job posting databases. We discuss a comprehensive methodology based on classic text mining techniques that discovers hidden connections between information generated by knowledge users and knowledge creators to find similarities between knowledge creators with knowledge users. The matching task is solved using a combination of unsupervised and supervised clustering and classification models. Experimental results on a number of performance metrics reveal the validity of both K-NN and SVM classifiers to classify this type of text records, with K-NN performing slightly better than SVM in a number of metrics (e.g., F1-score, Silhouette Validity Index).

This paper also contributes to the body of research by creating a methodology that can be adopted to create a system that solves similar problems. Finally, this paper illustrates its contribution by developing a prototype application that attempt to bridge the gap between research from industry and academia.

While presented in the context of matching knowledge users with knowledge creators, this solution can also be applied to similar problems like matching projects or tenders with contractors. In addition, more complex set of features that include citation, and co-citation data can be incorporated in future versions to enhance user experience and provide improved matching performance. Another way to enhance user experience is by incorporating more data sources. It would be interesting to find matches between patents, research publications and business organizations. It also would be interesting to find whether this prototype or similar ones can be reengineered to retrieve data not just by keyword search, but also by a particular research paper or job post, shifting the focus from a bunch of keywords to only one or a set of research papers or job posts. It would be interesting to see how the solution would perform under these conditions, and how matches are found and computed.

ACKNOWLEDGEMENTS

We are indebted to the Ministry of Higher Education in Saudi Arabia for their continued support.

REFERENCES

Antezana, E., Kuiper, M. and Mironov, V., 2009. Biological knowledge management: The emerging role of the Semantic Web technologies. Briefings in Bioinformatics, 10, pp.392–407.

Bielefeld University, 2014. About Bielefeld Academic Search Engine (BASE).

Bilenko, M. and Mooney, R.J., 2003. On evaluation and training-set construction for duplicate detection. In *Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation.* pp. 7–12.

Bozeman, B., 2000. Technology transfer and public policy: a review of research and theory. Research Policy, 29, pp.627–655.

Campbel, S. and Swigart, S., 2014. Go Beyond Google: *Gathering Internet Intelligence 5th editio., Cascade Insight.*

Carmel, D. et al., 2001. Static index pruning for information retrieval systems. In Proceedings of the *24th annual international ACM SIGIR conference on Research and development in information retrieval* - SIGIR '01. pp. 43–50.

Chitika Insights, 2013. Online Ad CTR: Impact of Referring Google Result Position

Chung, W., 2004. An automatic text mining framework for knowledge discovery on the Web. University of Arizona.

Cohen, W.W., 1998. Integration of heterogeneous databases without common domains using queries based on textual similarity. *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, 27, pp.201–212.

Colas, F. and Brazdil, P., 2006. Comparison of SVM and some older classification algorithms in text classification tasks. *IFIP International Federation for Information Processing*, 217, pp.169–178.

Council of Canadian Academies, 2012. The State of Science and Technology in Canada, Ottawa, Ontario.

Deerwester, S. et al., 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, pp.391–407.

Dooris, M.J., 1989. Organizational Adaptation and the Commercialization of Research Universities. Planning for Higher Education, 17(3), pp.21–31.

Dorneles, C.F., Gonçalves, R. and Santos Mello, R., 2010. Approximate data instance matching: a survey. Knowledge and Information Systems, 27(1), pp.1–21.

Elmagarmid, A.K., Ipeirotis, P.G. and Verykios, V.S., 2007. Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering, 19, pp.1–16.

Ertek, G., Tapucu, D. and Arin, I., 2013. Text mining with rapidminer. In M. Hofmann and R. Klinkenberg, eds. RapidMiner: *Data Mining Use Cases and Business Analytics Applications. Boca Raton, FL*: CRC Press, pp. 241–261.

Etzkowitz, H., 2002. Incubation of incubators: innovation as a triple helix of university–industry–government networks Henry. Science and Public Policy, 29, pp.115–128.

Etzkowitz, H. and Peters, L.S., 1991. Profiting from knowledge: Organisational innovations and the evolution of academic norms. Minerva, 29(2), pp.133–166.

Fellegi, I.P. and Sunter, A.B., 1969. A Theory for Record Linkage. Journal of the American Statistical Association, 64, pp.1183–1210.

Kannan, A. et al., 2011. Matching unstructured product offers to structured product specifications. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* - KDD '11. New York, New York, USA: ACM Press, pp. 404–412.

Karlsson, M., 2004. Commercialization of Research Results in the United States: *An Overview of Federal and Academic Technology Transfer*, Washington, DC.

Köpcke, H., Thor, A. and Rahm, E., 2010. Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, 3, pp.484–493.

Li, F. and Yang, Y., 2003. A Loss Function Analysis for Classification Categorization Methods in Text. In *Proceedings of the Twentieth International Conferenceon Machine Learning*. pp. 472–479.

Li, M., Li, H. and Zhou, Z.-H., 2009. Semi-supervised document retrieval. *Information Processing and Management*, 45(3), pp.341–355.

Liu, S.-H. et al., 2011. Development of a Patent Retrieval and Analysis Platform – A hybrid approach. Expert Systems with Applications, 38(6), pp.7864–7868.

Maedche, A. and Staab, S., 2001. Ontology learning for the Semantic Web. *IEEE Intelligent Systems*, 16(2), pp.72–79.

Mierswa, I. et al., 2006. YALE: Rapid prototyping for complex data mining tasks. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06. New York, New York, USA: ACM Press, pp. 935–940.

Mitkov, R., 2002. Anaphora Resolution 1st editio., New York, NY: Routledge.

Newcombe, H.B. et al., 1959. Automatic Linkage of Vital Records: Computers can be used to extract "follow-up" statistics of families from files of routine records. Science, 130(3381), pp.954–959.

Nidhi and Gupta, V., 2011. Recent Trends in Text Classification Techniques. International Journal of Computer Applications, 35(6), pp.45–51.

Nordfors, D., Sandred, J. and Wessner, C., 2003. Commercialization of Academic Research Results, Stockholm, Sweden: Swedish Agency for Innovation Systems.

Özgür, A., Özgür, L. and Güngör, T., 2005. Text Categorization with Class-Based and Corpus-Based Keyword Selection. In pInar Yolum et al., eds. *Proceedings of the 20th international conference on Computer and Information Sciences*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 606–615.

Pelleg, D. and Moore, A.W., 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *In Proceedings of the Seventeenth International Conference on Machine Learning*. pp. 727–734.

Porter, M.F., 1980. An algorithm for suffix stripping. Program: electronic library and information systems, 14(3), pp.130–137.

Ramesh, P., 2014. Prediction of cost overruns using ensemble methods in data mining and text mining algorithms. Rutgers, The State University of New Jersey.

Rogers, E.M., Takegami, S. and Yin, J., 2001. Lessons learned about technology transfer. *Technovation*, 21, pp.253–261.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics,* 20, pp.53–65.

Siegel, D.S. et al., 2004. Toward a model of the effective transfer of scientific knowledge from academicians to practitioners: qualitative evidence from the commercialization of university technologies. *Journal of Engineering and Technology Management*, 21(1-2), pp.115–142.

Sokolova, M., Japkowicz, N. and Szpakowicz, S., 2006. Beyond Accuracy , F-Score and ROC : A Family of Discriminant Measures for Performance Evaluation. In *Proceedings of the 19th Australian joint conference on Artificial Intelligence: advances in Artificial Intelligence*. pp. 1015–1021.

Swamidass, P.M. and Vulasa, V., 2009. Why university inventions rarely produce income? Bottlenecks in university technology transfer. *Journal of Technology Transfer*, 34, pp.343–363.

Winkler, W.E., 2002. Methods for Record Linkage and Bayesian Networks,

Xiang, G. et al., 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. *In Proceedings of the 21st ACM international conference on Information and knowledge management* - CIKM '12. p. 1980.

Yang, Y. and Liu, X., 1999. A re-examination of text categorization methods. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 99, pages, pp.42–49.

Zhou, L., Dai, L. and Zhang, D., 2007. Online shopping acceptance model – a critical survey of consumer factors in online shopping. *Journal of Electronic Commerce Research*, 8(1), pp.41–63.