Novi Sad J. Math.
Vol. 38, No. 3, 2008, 227-234

TEXT MINING: APPROACHES AND APPLICATIONS[1]

Miloš Radovanović,[2] Mirjana Ivanović[2]

Abstract. The field of text mining seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns. The techniques employed usually do not involve deep linguistic analysis or parsing, but rely on simple "bag-of-words" text representations based on vector space. Several approaches to the identification of patterns are discussed, including dimensionality reduction, automated classification and clustering. Pattern exploration is illustrated through two applications from our recent work: a classification-based Web meta-search engine and visualization of coauthorship relationships automatically extracted from a semi-structured collection of documents describing researchers in the region of Vojvodina. Finally, preliminary results concerning the application of dimensionality reduction techniques to problems in sentiment classification are presented.

AMS Mathematics Subject Classification (2000): 68U15, 68T50, 62H30

Key words and phrases: text mining, text categorization, clustering, dimensionality reduction, text visualization, sentiment classification

1. Introduction

Text mining is a new area of computer science which fosters strong connections with natural language processing, data mining, machine learning, information retrieval and knowledge management. Text mining seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns [2]. This paper will discuss several approaches to the identification of global patterns in text, based on the "bag-of-words" (BOW) representation described in Section 2. The covered approaches are automated classification and clustering (Section 3), and dimensionality reduction (Section 5). Pattern exploration will be illustrated through two applications from our recent work: presentation of Web meta-search engine results (Section 4) and visualization of coauthorship relationships automatically extracted from a semi-structured collection of documents describing researchers in the Serbian province of Vojvodina (Section 6). Finally, preliminary results concerning the application of dimensionality reduction techniques to problems in sentiment classification are presented in Section 7.

[1] This work was supported by project *Abstract Methods and Applications in Computer Science* (no. 144017A), of the Serbian Ministry of Science and Environmental Protection

[2] University of Novi Sad, Faculty of Science, Department of Mathematics and Informatics, Trg D. Obradovića 4, 21000 Novi Sad, Serbia, e-mail: {radacha, mira}@dmi.uns.ac.rs

2. Bag-of-Words Document Representation

Let W be the *dictionary* – the set of all terms (words) that occur at least once in a collection of documents D. The bag-of-words representation of document d_n is a vector of weights $(w_{1n},\ldots,w_{|\mathrm{W}|n})$. In the simplest case, the weights $w_{in} \in \{0,1\}$ and denote the presence or absence of a particular term in a document. More commonly, w_{in} represent the frequency of the ith term in the nth document, resulting in the term frequency representation. Normalization can be employed to scale the term frequencies to values between 0 and 1, accounting for differences in the lengths of documents. Besides words, *n-grams* may also be used as terms. However, two different notions have been referred to as "n-grams" in the literature. The first are *phrases* as sequences of n words, while the other notion are n-grams as sequences of *characters*. N-grams as phrases are usually used to *enrich* the BOW representation rather than on their own. N-grams as sequences of characters are used *instead* of words.

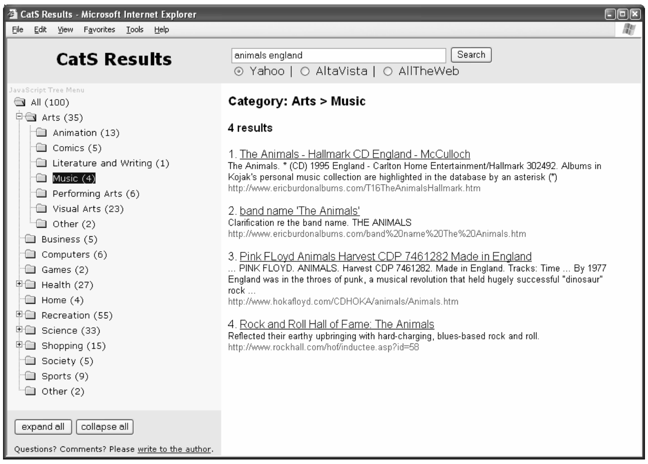
The transformation of a document set D into the BOW representation enables the transformed set to be viewed as a matrix, where rows represent document vectors, and columns are terms. This view enables various matrix decomposition techniques to be applied for the tasks of clustering [2] and dimensionality reduction (Section 5). Furthermore, since documents are treated as vectors, they can be compared using classical distance/similarity measures. The most commonly employed measures include cosine and Tanimoto similarity [6].

3. Machine Learning with Textual Data

The field of *machine learning* (ML) is concerned with the question of how to construct computer programs that automatically improve with experience. One important division of learning methods is into *supervised* and *unsupervised*. In supervised learning, computer programs capture structural information and derive conclusions (predictions) from previously labeled *examples* (instances, points). Unsupervised learning finds groups in data without relying on labels.

ML techniques can roughly be divided into four distinct areas: classification, clustering, association learning and numeric prediction [10]. Classification applied to text is the subject of *text categorization* (TC), which is the task of automatically sorting a set of documents into *categories* from a predefined set [8]. Classification of documents is employed in text filtering, categorization of Web pages (see Section 4), sentiment analysis (see Section 7), etc. Classification can also be used on smaller parts of text depending on the concrete application, e.g. document segmentation or topic tracking. In the ML approach, classifiers are trained beforehand on previously sorted (labeled) data, before being applied to sorting unseen texts. The most popular classifiers applied to text include naïve Bayes, k-nearest neighbor, and support vector machines [10].

While classification is concerned with finding models by *generalization* of evidence produced by a dataset, clustering deals with the *discovery* of models by finding groups of data points which satisfy some objective criterion, e.g. maximize inter-cluster similarity of points, while minimizing similarity of points from



Category: Arts > Music

4 results

1. The Animals - Hallmark CD England - McCulloch
The Animals. * (CD) 1995 England - Carlton Home Entertainment/Hallmark 302492. Albums in Kojak's personal music collection are highlighted in the database by an asterisk (*)
http://www.ericburdonalbums.com/T16TheAnimalsHallmark.htm

2. band name 'The Animals'
Clarification re the band name. THE ANIMALS
http://www.ericburdonalbums.com/band%20name%20The%20Animals.htm

3. Pink FLoyd Animals Harvest CDP 7461282 Made in England
... PINK FLOYD. ANIMALS. Harvest CDP 7461282. Made in England. Tracks: Time ... By 1977 England was in the throes of punk, a musical revolution that held hugely successful "dinosaur" rock ...
http://www.hokafloyd.com/CDHOKA/animals/Animals.htm

4. Rock and Roll Hall of Fame: The Animals
Reflected their earthy upbringing with hard-charging, blues-based rock and roll.
http://www.rockhall.com/hof/inductee.asp?id=58

Figure 1: Results for query 'animals england' classified into *Arts* → *Music*

different clusters. Examples of algorithms used on text data include k-means, and approaches employing matrix decompositions [2].

4. Application: Enhancing Web Search

One way to enhance users' efficiency and experience of Web search is by means of *meta-search engines*. Traditionally, meta-search engines were conceived to address different issues concerning general-purpose search engines, including Web coverage, search result relevance, and their presentation to the user. A common approach to alternative presentation of results is by sorting them into (a hierarchy of) clusters which may be displayed to the user in a variety of ways, e.g. as a separate expandable tree (vivisimo.com) or arcs which connect Web pages within graphically rendered "maps" (kartoo.com). However, topics generated by clustering may not prove satisfactory for every query, and the "silver bullet" method has not yet been found. An example of a meta-search engine which sorts search results into a hierarchy of topics using *text categorization* techniques is CatS [7] (stribog.im.ns.ac.yu/cats). Figure 1 shows the subset of the 100 results for query 'animals england' sorted into category *Arts* → *Music*, helping separate pages about animals living in England from pages concerning the English music scene. The categories employed by CatS were extracted from the dmoz Open Directory (www.dmoz.org).

5. Dimensionality Reduction

It is clear that even for small document collections the BOW document vector will have high dimensionality. This may hinder the application of ML methods not only for technical reasons, but also by degrading the performance of learning algorithms which cannot scale to such high dimensions. There are two approaches to the problem of dimensionality reduction: feature *selection*, where the resulting set of features is a subset of the old one, and feature *extraction*, which derives a different (but smaller) set of features. Feature selection and extraction techniques can be divided into *supervised* and *unsupervised* methods. As for feature extraction, another division is into *linear* and *nonlinear* techniques. This paper will focus on linear methods only.

Linear feature extraction techniques seek to find a transformation matrix G, which, when multiplied with the document-term matrix A, gives a matrix representing documents in a new space: $A' = AG$. The number of columns of G can be significantly smaller than the number of columns of A, producing a representation of documents in A' of much lower dimensionality. The transformation matrix G is derived from the document-term matrix A, but may be applied in the same way to (matrices of row) vectors which were not used in the derivation.

One of the most notable feature extraction methods used on text is singular value decomposition (SVD). The SVD of matrix $A = U\Sigma V^T$, where the columns of U are orthogonal eigenvectors of AA^T, the columns of V are orthogonal eigenvectors of $A^T A$, and Σ is a diagonal matrix of singular values – the square roots of eigenvalues of AA^T. For SVD reduction, matrix G consists of columns of V that correspond to largest singular values. Application of SVD to BOW data has an intuitive interpretation – the new features correspond to combinations of original terms and represent semantic "concepts" ("topics") that were derived from co-occurrence relations of terms in different documents.

Another useful technique for linear feature extraction is multidimensional scaling (MDS), which projects high dimensional vectors to a lower dimensional space at the same time trying to preserve pairwise distances between points. MDS is especially suitable for visualization of a collection of documents, which will be exemplified in Section 6.

When class information is available for a document set, supervised linear feature extraction methods may be used to reduce dimensionality, as well as increase the separation between classes. One classic example is linear discriminant analysis (LDA), which produces a linear transformation of data that attempts to minimize within-class scatter (a measure which features covariances of data points belonging to the same class) and maximize between-class scatter (featuring covariances of class means). LDA is usually applied to already reduced dense matrices, for example with SVD [9]. Another well-known example of an approach to supervised linear feature extraction is partial least squares (PLS), more precisely the efficient SIMPLS algorithm [3], which tries to maximize the covariance between vectors of the new representation and output vectors.

6. Application: Mining Bibliographic Data

Vojvodina, the northern province of Serbia, is home to many educational and research institutions. In 2004, the Provincial Secretariat for Science and Technological Development of Vojvodina started collecting data from researchers employed at institutions within its jurisdiction. Every researcher was asked to

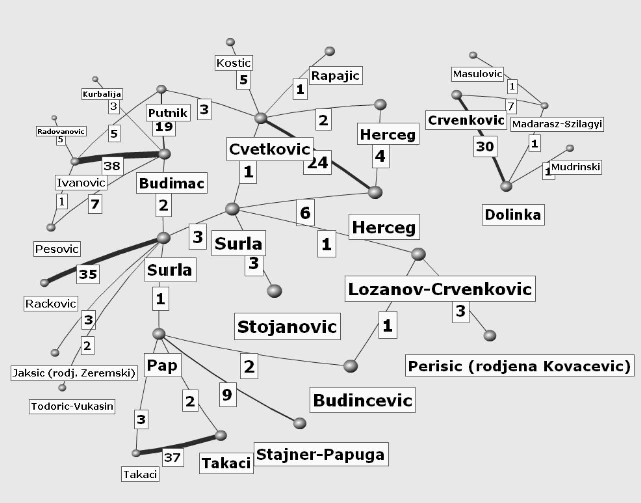


Figure 2: Collaboration diagram of researchers from the DMI

fill in a form, provided as an MS Word document, with bibliographic references of all authored publications, among other data. Notable properties of the collection are its incompleteness and diversity of approaches to giving references, permitted by information being entered in free text format.

The work outlined in this section consists of three main tasks: (1) extract basic data about organizations, researchers, and full lists of bibliographic references from the collection, (2) detect duplicate references in forms corresponding to different researchers, thus establishing coauthorship relations, and (3) import the extracted data to IST World Web portal and use its visualization functionalities for analysis of the extracted information [6].

The collection of documents dated July 6, 2006 includes 2,278 researchers from 60 institutions. We programmed an extractor which is able to automatically isolate every researcher's data and save it as CERIF compliant XML to enable quick import into the IST World relational database. Furthermore, the extractor compares references, detecting coauthorships between researchers. In summary, the extractor isolates 110,394 references and detects 30,822 duplicates, making the total number of references in the database 79,572.

To calculate the similarity of two references the extractor uses Tanimoto similarity on character 2-grams. Detected coauthorships enable the visualization of collaboration between researchers and organizations in IST World. Figure 2 depicts the collaboration of researchers from the Department of Mathematics and Informatics (DMI) of the Faculty of Science, University of Novi Sad.

IST World constructs a competence diagram of researchers by merging the titles of their references into a single document for each researcher, and applying SVD and MDS to construct a two-dimensional visual representation. The competence diagram for researchers from the DMI is presented in Fig. 3.

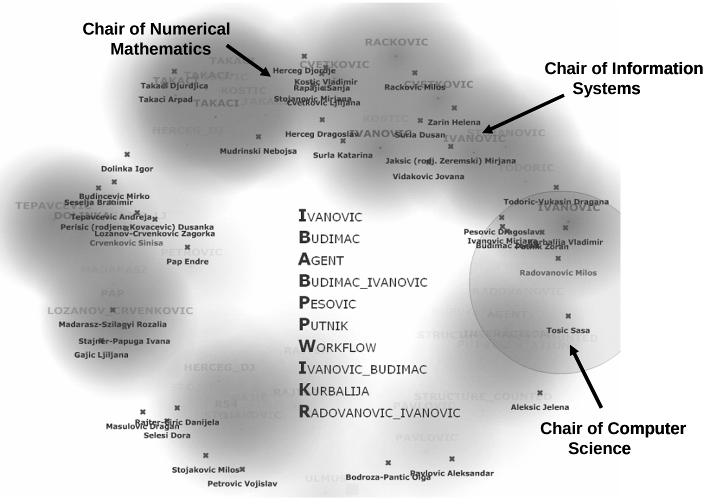


Figure 3: The competence diagram of researchers from the DMI

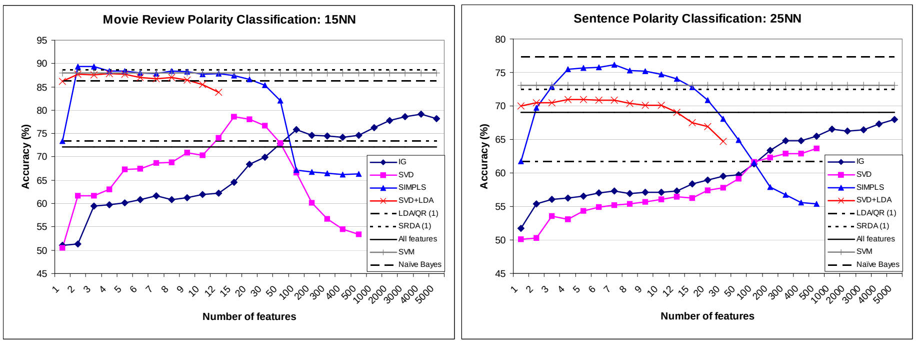
7. Application: Sentiment Classification

The field of sentiment analysis deals with the analysis of opinions found in documents. One of the basic tasks is sentiment *classification*, where units of text are sorted into classes which correspond to, e.g. positivity or negativity of expressed opinion. This section will present an empirical study into the application of dimensionality reduction techniques to two sentiment classification problems: (1) document polarity classification, where documents representing complete reviews are classified into positive or negative, and (2) sentence polarity classification, which deals with polarity classification of individual sentences.

For document polarity classification the movie review dataset from [4] was used. It consists of 1000 positive and 1000 negative reviews collected from www.rottentomatoes.com. The sentence polarity dataset is comprised of 10,662 movie review snippets from the same source, with each snippet labeled by its source review's class [5]. We used a normalized term frequency BOW representation which includes words, 2-grams and 3-grams. The number of terms is 69,459 for movie review data, and 14,478 for the sentence polarity dataset.

Dimensionality reduction methods included in the study are information gain (IG) [10], SVD, SIMPLS, and various variants of LDA: LDA/QR which uses QR matrix decomposition [11], and SRDA which takes a spectral regression approach [1]. Although more efficient than classic LDA, the two methods still have the limitation that the number of features in the new space has to be less than the number of classes, limiting the number of features in the reduced space for the considered problems to one. We therefore also considered an approach that applies SVD to reduce the datasets to 500 dimensions, and then an LDA method which can produce more features than there are classes [9].

The results of the evaluation of 15- and 25-nearest neighbor classifier on the



(a) (b)

Figure 4: Ten-fold cross-validation accuracy of nearest neighbor classification and various dimensionality reduction methods on (a) movie review polarity data, and (b) sentence polarity data

two datasets are shown in Fig. 4. Accuracy is averaged over 10 cross-validation folds, with the folds of the movie review data being the same as those used in [4]. It is evident from the charts that supervised feature extraction methods are able to greatly reduce the dimensionality of data, at the same time improving classification accuracy of the baseline nearest neighbor classifier, and also beat or be comparable to naïve Bayes and SVM evaluated on full feature sets. The best reduction method on both datasets is SIMPLS, which permits the reduction of dimensionality to less than 10 features while obtaining excellent accuracy for both problems. In an online application setting this would enable the use of indexing techniques for quick retrieval of nearest neighbors and prediction of the polarity of a new review, as well as efficient addition of new reviews to the predictor, with periodic offline recalculation of the SIMPLS reduction.

Although supervised feature extraction techniques have already been applied to text categorization problems [9, 12], the considered problems were in the realm of *topical* text classification, with a much larger number of classes. In the sentiment classification problems considered here, supervised feature extraction techniques were effective at a much lower number of features. Furthermore, observing the performance of kNN on topical classification in [12] reveals that the difference between PLS and SVD is more pronounced in polarity classification problems. SVD is known to produce features that correspond to latent topics – such features are suitable for topical classification, but apparently not so much for polarity classification. According to the presented evidence, supervision within feature extraction techniques can be particularly useful for distinguishing differences in polarity.

8. Conclusion

Classification, clustering and dimensionality reduction are only some of the methods; enhancing Web search, mining bibliographic data and sentiment classification are only some of the applications. The aim of this paper was to provide an illustration of the already vast field of text mining through discussion of our recent work, together with preliminary results which connect dimensionality reduction with sentiment classification.

References

[1] Cai, D., He, X., Han, J., SRDA: An efficient algorithm for large-scale discriminant analysis. IEEE T. Knowl. Data En., 20(1) (2008), 1–12.

[2] Feldman, R., Sanger, J., The Text Mining Handbook. Cambridge University Press, 2007.

[3] de Jong, S., SIMPLS: An alternative approach to partial least squares regression. Chemometr. Intell. Lab., 18(3) (1993), 251–263.

[4] Pang, B., Lee, L., A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the ACL, pages 271–278, 2004.

[5] Pang, B., Lee, L., Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the ACL, pages 115–124, 2005.

[6] Radovanović, M., Ferlež, J., Mladenić, D., Grobelnik, M., Ivanović, M., Mining and visualizing scientific publication data from Vojvodina. Novi Sad Journal of Mathematics 37(2) (2007), 161–180.

[7] Radovanović M., Ivanović, M., CatS: A classification-powered meta-search engine. In: Last, M., et al., editors, Advances in Web Intelligence and Data Mining, pages 191–200, Springer-Verlag, 2006.

[8] Sebastiani, F., Text categorization. In Zanasi, A., editor, Text Mining and its Applications, pages 109–129, Southampton, UK: WIT Press, 2005.

[9] Torkkola K., Linear discriminant analysis in document classification. In: IEEE ICDM Workshop on Text Mining, pages 800–806, 2001.

[10] Witten, I.H., Frank, E., Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, 2nd edition, 2005.

[11] Ye, J., Li, Q., A two-stage linear discriminant analysis via QR-decomposition. IEEE T. Pattern Anal. 27(6) (2005), 929–941.

[12] Zeng, X.-Q., Wang, M.-W., Nie, J.-Y., Text classification based on partial least square analysis. In: Proceedings of ACM SAC, pages 834–838, 2007.

Received by the editors October 31, 2008