Inf Retrieval (2011) 14:208–211
DOI 10.1007/s10791-010-9153-5

BOOK REVIEW

Michael W. Berry and Jacob Kogan (eds.): Text mining: applications and theory

John Wiley and Sons, Ltd, 2010, xiv + 207 pp, £55.00/€66.00, hardcover, ISBN: 978-0-470-74982-1

Zhang Xiaojun

Published online: 23 November 2010

Text mining: *applications and theory* presents the state-of-the-art algorithms for text mining from both the academic and industrial perspectives. Actually, it is the proceedings of a one-day workshop on text mining which hold on May 2, 2009 in conjunction with the SIAM Ninth International Conference on Data Mining. This volume demonstrates how advancements in the fields of applied mathematics, computer science, machine learning, and natural language processing can collectively capture, classify, and interpret words and their contexts. As suggested in the preface, text mining is needed when "words are not enough" (p. xiv). Collectively, the contributors span several major topic areas in text mining: Keyword extraction, Classification and clustering, Anomaly and trend detection, Text streams. According to these topic areas, Michael W. Berry and Jacob Kogan divided the content into three parts: Part I, Text Extraction, Classification and Clustering (Chaps. 1–5); Part II, Anomaly and Trend Detection (Chaps. 6–8); and Part III, Text Streams (Chaps. 8–9).

In Chap. 1, "Automatic keyword extraction from individual documents", the contributors describe Rapid Automatic Keyword Extraction (RAKE), an unsupervised, domain-independent, and language-independent method for extracting keywords from individual documents. Their presented results on a benchmark dataset of technical abstracts show that RAKE is more computationally efficient than Text Rank while achieving higher precision and comparable recall scores. Then they describe a novel method for generating stop lists, which we use to configure RAKE for specific domains and corpora. Finally, they apply RAKE to a corpus of news articles and define metrics for evaluating the exclusivity, essentiality, and generality of extracted keywords, enabling a system to identify keywords that are essential or general to documents in the absence of manual annotations. Ideally, keywords represent in condensed form the essential content of a document, and keywords are widely used to define queries within information retrieval (IR) systems as they are easy to define, revise, remember, and share. Therefore, a relatively high rate of keywords extraction precision should be vital important to the IR systems. Unfortunately, the

Z. Xiaojun (✉)
School of Foreign Languages, Shaanxi Normal University,
199 South Chang'an Road, 710062 Xi'an, China
e-mail: Andy_zxj@126.com

presented result of this chapter shows that the precision is only 67% even though the three false positives were revealed to be the main factors to cause the result.

In Chap. 2, "Algebraic techniques for multilingual document clustering", it reviews a number of computational techniques for clustering documents in a multilingual corpus, provides some additional insight into these techniques, and presents some recent advances. Specifically, it show multiple algebraic models that were developed recently and that use matrix and tensor manipulations. These models include the singular value decomposition (SVD) or latent semantic analysis (LSA), Tucker1, PARAFAC2, LSA with term alignments (LSATA), Latent morpho-semantic analysis (LMSA) and LMSA with term alignment (LMSATA). Based on the aggregate results of all these algebraic models in average P1 (precision at one document) and average MP5 (multilingual precise at five documents) scores, it concludes that (1) the SVD-based techniques, such as LSA and LMSA, are the fastest; (2) the eigenvector-based approach of LSATA and LMSATA requires more time due to the larger matrix and term-alignment step; and (3) the tensor-based techniques Tucker1 and PARAFAC2 are the slowest due to the data being organized as a large three-way array. It recommends the best method reported in this chapter, LMSATA. To this "best" choice, we need a broad collection of techniques to pre-process including: (1) morphological analysis of language using techniques from statistical machine translation; (2) techniques from latent semantic analysis, including dimentionality reduction using the SVD; and (3) numerical linear algebra for simultaneously analyzing term co-occurrences and term-term alignments. You known, it is quite difficult to do so. Meanwhile, this chapter just compared the results of the above SVD-based models, and the traditional vector space models (SVM) were ignored. Anyway, a good comment to this chapter is that these methods can be applied not just to pairs of languages, but also to groups of languages when a suitable multi-parallel corpus exists.

In Chap. 3, "Content-based spam email classification using machine-learning algorithms", the authors consider five supervised machine-learning algorithms for an evaluation study of spam filtering application. The algorithms selected in this study include: naïve Bayes classifier (NB), support vector machines (SVMs), logitBoost algorithm (LB), augmented latent semantic indexing space model (LSI) and radial basis function (RBF) networks. In this study, two benchmark email testing corpora are selected for experiments that were constructed from two different languages and have reverse ratios of the number of spam emails to analyze the usefulness of feature selection for these algorithms. The experiment results show that, in terms of adaptability to cost-sensitive spam filtering, the classifiers based on LSI and RBF demonstrate their strength in this evaluation. The both competitive alternatives share a common characteristic that a clustering component is definitely utilized in their model training.

Another email classification study in Chap. 4, "Utilizing nonnegative matrix factorization for email classification problems", introduces an approximation technique, the nonnegative matrix factorization (NMF), and its application to the task of email classification. NMF demines reduced rank nonnegative factors W (interpreted as basis *email* or the basis *features*) and H (interpreted as basis coefficients) which approximate a given nonnegative data matrix A, such that $A \approx WH$. If NMF is applied to an *email* $\times$ *feature* matrix, then W contains k basis *features*. Therefore, NMF algorithm is a k-means technique and $k = 3$ in this study which correspond to ham, spam and phishing three basis email messages. Although it explained the feature subset selection methods, *gain* and *gain ratio*, the three features are still veiled in this study. What are they?

As to k-means algorithms, Chap. 5, "Constrained clustering with k-means type algorithms" focuses on three k-means type clustering algorithms and two different distance-like

functions. The clustering algorithms are *k*-means, smoka, and spherical *k*-means. And the distant-like functions are "reverse Bregman divergence" and 'cosine similarity'. The study shows that these algorithms and distance-like functions can be reduced to clustering with cannot-link constraints only. Therefore the authors substitute cannot-link constraints by penalty, and propose clustering algorithms that tackle clustering with penalties so that each algorithm is capable of clustering a vector dataset equipped with must-link constraints and a penalty function that penalized violations of cannot-link constraints. Numerical experiments with *k*-means and spherical *k*-means algorithms show improvement of clustering performance in the presence of constraints while the experiments result of smoka "will be reported elsewhere" (p. 102).

Compared to books on the same theme (Berry 2003; Weiss et al. 2005; Srivastava et al. 2009), the second part should be something really new. Chap. 6 begins with the Part II of this book. In this chapter entitled "Survey of text visualization techniques", the contributors explore several visual techniques and describe specific examples of software that utilizes them. On the perspective of user's needs or purposes, the study introduce, respectively, variations of the time line plot techniques, tag clouds and other similar techniques, and sentiment tracking and its related visualization software, respectively. One of the visual post-processing tools which tailored for specific text mining packages, entitled FutureLens, is also discussed in great detail in this chapter.

Chapter 7, "Adaptive threshold setting for novelty mining", addresses the problem of setting an adaptive threshold by utilizing user feedback over time. The proposed method, the Gaussian-based adaptive threshold setting (GATS) algorithm, modeled the distributions of novelty scores from both novel and nonnovel classes by the Gaussian distributions. GATS is a general method, which can be tuned according to different performance requirements, by combining with different optimization criteria. In this chapter, the most commonly used performance evaluation measure in novelty mining, the F score, has been employed as the optimization criteria. The experimental results suggest that GATS is able to meet the different performance requirements by setting the weights of precision and recall externally.

Chapter 8, "Text mining and cybercrime", is the end of Part II. Cyberbullying and Internet predation frequently occur over an extended period of time and across several technological platforms (i.e. chat rooms, social networking sites, cell phones, etc.). In this chapter, it describes the current state of research in the areas of cyberbullying and Internet predation, introduces several commercial products which claim to provide chat and social networking site monitoring for home use, and discusses opportunities for future research into this interesting and timely field.

There are only two chapters included in the last part of this book. Chapter 9, "Events and trends in text streams", targets at developing algorithms to find and characterize in topic within text stream. Specifically, it has implemented some of these algorithms into a *surprise* event and *emerging* trend detection designed to monitor a stream of text or messages for changes within the content of that data stream. Terminologically, it refers to the instantaneous discontinuity types (point discontinuity or jump discontinuity) as *surprise* event and a change in topic for extended period of time (jump discontinuity or slope discontinuity) as *emerging* trend. The term "*surprise*" is used a little puzzled for the authors name it as "*surprising*" sometimes.

The last chapter, "Embedding semantics in LDA topic models", specifies a statistical sampling technique to describe how words in documents are generated based on a small set of hidden topics. In this chapter, the contributors investigate the role of prior knowledge semantics in estimating the topical structure of large text data in both batch and online

modes under the framework of latent Dirichlet alglocation (LDA) topic modeling. The objective is to enhance the descriptive and/or predictive model of the data's thematic structure based on the embedded prior knowledge about the domain's semantics.

Generally, the articles in the first part are all technical reports and most of the ones in the second and third part are field reviews.

In general, *Text mining*: *application and theory* provides state-of-the-art algorithms and techniques for critical tasks in text mining applications, such as clustering, classification, anomaly and trend detection, and stream analysis, presents a survey of text visualization techniques and looks at the multilingual text classification problem, discusses the issue of cybercrime associated with chatrooms, features advances in visual analytics and machine learning along with illustrative examples, and, is accompanied by a supporting website (http://www.wiley.com/go/berry_mining) featuring datasets. It is extremely useful for practitioners and students in computer science, natural language processing, bioinformatics and engineering who wish to use text mining techniques.

References

Berry, M. W. (Ed.). (2003). *Survey of text mining*. New York, USA: Springer.
Srivastava, A. N., & Sahami, M. (Eds.). (2009). *Text mining*. Boca Raton, USA: CRC Press.
Weiss, S. M., et al. (Eds.). (2005). *Text mining*. New York, USA: Springer.