Koessler, Denise R.; **Martin, Benjamin W.**; **Kiefer, Bruce E.**; **Berry, Michael W.**
The effects of tabular-based content extraction on patent document clustering. (English)
Zbl 07042137
Algorithms (Basel) 5, No. 4, 490-505 (2012).

Summary: Data can be represented in many different ways within a particular document or set of documents. Hence, attempts to automatically process the relationships between documents or determine the relevance of certain document objects can be problematic. In this study, we have developed software to automatically catalog objects contained in HTML files for patents granted by the United States Patent and Trademark Office (USPTO). Once these objects are recognized, the software creates metadata that assigns a data type to each document object. Such metadata can be easily processed and analyzed for subsequent text mining tasks. Specifically, document similarity and clustering techniques were applied to a subset of the USPTO document collection. Although our preliminary results demonstrate that tables and numerical data do not provide quantifiable value to a document's content, the stage for future work in measuring the importance of document objects within a large corpus has been set.

MSC:

68 Computer science
92 Biology and other natural sciences

Keywords:

text mining; patent documents; table data

Software:

PERL; TMG

Full Text: DOI

References:

[1] Kochi, T.; Saitoh, T.; A layout-free method for extracting elements from document images; Document Analysis Systems: Theory and Practice: Berlin, Germany 1999; Volume Volume 1655 ,215-224.

[2] Gupta, G.; Niranjan, S.; Shrivastava, A.; Sinha, R.; Document layout analysis and classification and its application in OCR; EDOCW '06. 10th IEEE International, Proceedings of Enterprise Distributed Object Computing Conference Workshops: ; ,58-58.

[3] Sharpe, M.; Ahmed, N.; Sutcliffe, G.; An intelligent document understanding \& reproduction system; MVA: 1994; Volume 267 ,267-271.

[4] Berry, M.; Esau, R.; Kiefer, B.; The use of text mining techniques in electronic discovery for legal matters; Next Generation Search Engines: Advanced Models for Information Retrieval: Hershey, PA, USA 2012; ,174-190.

[5] Vincent, L.; Google book search: document understanding on a massive scale; Proceedings of International Conference on Document Analysis and Recognition: Los Alamitos, CA, USA 2007; Volume Volume 2 ,819-823.

[6] Yoon, B.; Park, Y.; A text-mining-based patent network: Analytical tool for high-technology trend; J. High Technol. Manag. Res.: 2004; Volume 15 ,37-50.

[7] Tseng, Y.H.; Lin, C.J.; Lin, Y.I.; Text mining techniques for patent analysis; Inf. Process. \& Manag.: 2007; Volume 43 ,1216-1247.

[8] Fattori, M.; Pedrazzi, G.; Turra, R.; Text mining applied to patent mapping: A practical business case; World Pat. Inf.: 2003; Volume 25 ,335-342.

[9] Farrow, G.S.D.; Xydeas, C.S.; Oakley, J.P.; Khorabi, A.; Prelcic, N.G.; A comparison of system architectures for intelligent document understanding; Signal Process.: Image Commun.: 1996; Volume 9 ,1-19.

[10] Malerba, D.; Ceci, M.; Berardi, M.; Machine learning for reading order detection in document image understanding; Machine Learning in Document Analysis and Recognition: Berlin, Germany 2008; ,45-70.

[11] Michel, J.; Bettels, B.; Patent citation analysis. A closer look at the basic input data from patent search reports; Scientometrics: 2001; Volume 51 ,185-201.

[12] Berry, M.; Drmac, Z.; Jessup, E.; Matrices, vector spaces, and information retrieval; SIAM Review: 1999; Volume 41 ,335-362. · Zbl 0924.68069

[13] Robertson, S.; Understanding inverse document frequency: On theoretical arguments for IDF; J. Doc.: 2004; Volume 60 ,503-520.

[14] Papineni, K.; Why inverse document frequency?; Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies: Stroudsburg, PA, USA 2001; ,25-32.

[15] Zeimpekis, D.; Gallopoulos, E.; TMG: A MATLAB toolbox for generating term-document matrices from text collections; Grouping Multidimensional Data: Recent Advances in Clustering: Berlin, Germany 2006; ,187-210.

[16] The Perl Programming Language Home Page; ; .

[17] MathWorks Home Page; ; .

[18] Cornell SMART Project English Stoplist; ; .