de Castro, Pablo A. D.; **de França, Fabrício O.**; **Ferreira, Hamilton M.**; **Palermo Coelho, Guilherme**; **Von Zuben, Fernando J.**

Query expansion using an immune-inspired biclustering algorithm. (English) Zbl 1205.68143

Nat. Comput. 9, No. 3, 579-602 (2010).

Summary: Query expansion is a technique utilized to improve the performance of information retrieval systems by automatically adding related terms to the initial query. These additional terms can be obtained from documents stored in a database. Usually, this task is performed by clustering the documents and then extracting representative terms from the clusters. Afterwards, a new search is performed in the whole database using the expanded set of terms. Recently, the authors have proposed an immune-inspired algorithm, namely BIC-aiNet, to perform biclustering of texts.

Biclustering differs from standard clustering algorithms in the sense that the former can detect partial similarities in the attributes. The preliminary results indicated that our proposal is able to group similar texts effectively and the generated biclusters consistently presented relevant words to represent a category of texts. Motivated by this promising scenario, this paper better formalizes the proposal and investigates the usefulness of the whole methodology on larger datasets. The BIC-aiNet was applied to a set of documents aiming at identifying the set of relevant terms associated with each bicluster, giving rise to a query expansion tool. The obtained results were compared with those produced by two alternative proposals in the literature, and they indicate that these techniques tend to generate complementary results, as a consequence of the use of distinct similarity metrics.

MSC:

68P15 Database theory
68P20 Information storage and retrieval of data

Keywords:

biclustering; artificial immune systems; information retrieval; query expansion

Software:

TMG

Full Text: DOI

References:

[1] Ada GL, Nossal GJV (1987) The clonal selection theory. Sci Am 257:50–57 · doi:10.1038/scientificamerican0887-62

[2] Agrawal R, Gehrke J, Gunopulus D, Raghavan P (1998) Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of the ACM/SIGMOD international conference on management of data, pp 94–105

[3] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403:503–510. doi: 10.1038/35000501 · doi:10.1038/35000501

[4] Castro PAD, de França FO, Ferreira HM, Von Zuben FJ (2007a) Applying Biclustering to Perform Collaborative Filtering. In: Proc. of the 7th International Conference on Intelligent Systems Design and Applications (ISDA), p 421–426, Brazil

[5] Castro PAD, de França FO, Ferreira HM, Von Zuben FJ (2007b) Evaluating the Performance of a Biclustering Algorithm Applied to Collaborative Filtering–A Comparative Analysis. In: Proc. of the 7th International Conference on Hybrid Intelligent Systems (HIS), p 65–70, Germany

[6] Castro PAD, de França FO, Ferreira HM, Von Zuben FJ (2007c) Applying biclustering to text mining: an immune-inspired approach. In: Proceedings of the 6th international conference on artificial immune systems (ICARIS), Brazil, pp 83–94

[7] Cheng Y, Church GM (2000) Biclustering of expression data. In: Proceedings of the 8th international conference on intelligent systems for molecular biology, pp 93–103

[8] Cho R, Campbell M, Winzeler E, Steinmetz L, Conway A, Wodicka L, Wolfsberg T, Gabrielian A, Landsman D, Lockhart

D, Davis R (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 2:65–73. doi: 10.1016/S1097-2765(00)80114-8 · doi:10.1016/S1097-2765(00)80114-8

[9] Coelho GP, de França FO, Von Zuben FJ (2008) A multi-objective multipopulation approach for biclustering. In: Proceedings of 7th international conference on artificial immune systems (ICARIS), vol 5132, pp 71–82

[10] Croft WB, Cook R, Wilder D (1995) Providing government information on the Internet: experiences with THOMAS. In: Proceedings of the 2nd international conference on the theory and practice of digital libraries, pp 19–24

[11] de Castro LN, Timmis J (2002) Artificial immune systems: a new computational intelligence approach. Springer Verlag, London · Zbl 1027.68108

[12] de Castro LN, Von Zuben FJ (2001) aiNet: an artificial immune network for data analysis. In: Data mining: a heuristic approach, pp 231–259

[13] de Castro LN, Von Zuben FJ (2002) Learning and optimization using the clonal selection principle. IEEE Trans Evol Comput 6(3):239–251. doi: 10.1109/TEVC.2002.1011539 · Zbl 05451946 · doi:10.1109/TEVC.2002.1011539

[14] Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6(2):182–197. doi: 10.1109/4235.996017 · Zbl 05451853 · doi:10.1109/4235.996017

[15] Dhillon IS (2001) Co-clustering documents and words using bipartite spectral graph partitioning. In: Proceedings of the 7th international conference on knowledge discovery and data mining, pp 269–274

[16] Divina F, Aguilar–Ruiz JS (2007). A multi-objective approach to discover biclusters in microarray data. In: Proceedings of the genetic and evolutionary computation conference (GECCO'07), London, UK,pp 385–392

[17] Feldman R, Sanger J (2006) The Text Mining Handbook. Cambridge University Press

[18] Giráldez R, Divina F, Pontes B, Aguilar–Ruiz JS (2007). Evolutionary search of biclusters by minimal intrafluctuation. In Proceedings of the IEEE international fuzzy systems conference (FUZZ–IEEE 2007), London, UK, pp 1–6

[19] Hartigan JA (1972) Direct clustering of a data matrix. J Am Stat Assoc 67(337):123–129. doi: 10.2307/2284710 JASA · doi:10.2307/2284710

[20] Jerne NK (1974) Towards a network theory of the immune system. Ann Immunol (Inst Pasteur) 125C:373–389

[21] Jones KS (1972) A statistical interpretation of term specificity and its application in retrieval. J Doc 28(1):11–21. doi: 10.1108/eb026526 · doi:10.1108/eb026526

[22] Lang K (1995) Newsweeder: learning to filter netnews. In: Proceedings of the twelfth international conference on machine learning, pp 331–339

[23] Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinform 1:24–25. doi: 10.1109/TCBB.2004.2 · Zbl 05103330 · doi:10.1109/TCBB.2004.2

[24] Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge · Zbl 1160.68008

[25] Maulik U, Mukhopadhyay A, Bandyopadhyay S, Zhang MQ, Zhang X (2008). Multiobjective fuzzy biclustering in microarray data: method and a new performance measure. In: Proceedings of the 2008 IEEE congress on evolutionary computation (CEC 2008), Hong Kong, China, pp 1536–1543

[26] Mitra S, Banka H (2006) Multi-objective evolutionary biclustering of gene expression data. Pattern Recognit 39:2464–2477. doi: 10.1016/j.patcog.2006.03.003 · Zbl 1103.68775 · doi:10.1016/j.patcog.2006.03.003

[27] Mitra S, Banka H, Pal SK (2006). A more framework for biclustering of microarray data. In: Proceedings of the 18th international conference on pattern recognition (ICPR'06), Hong Kong, China, pp 1154–1157

[28] Sheng Q, Moreau Y, De Moor B (2003) Biclustering microarray data by Gibbs sampling. Bioinformatics 19(2):196–205. doi: 10.1093/bioinformatics/btg1078 · doi:10.1093/bioinformatics/btg1078

[29] Symeonidis P, Nanopoulos A, Papadopoulos A, Manolopoulos Y (2007). Nearestbiclusters collaborative filtering with constant values. In: Advances in web mining and web usage analysis, vol 4811, Lecture notes in computer science. Springer-Verlag, Philadelphia, pp 36–55

[30] Tanay A, Sharan R, Shamir R (2005) Biclustering algorithms: a survey. In: Aluru S (ed) Handbook of computational molecular biology. Chapman \& Hall/CRC Computer and Information Science Series, Boca Raton, FL

[31] Tang C, Zhang L, Zhang I, Ramanathan M (2001) Interrelated two-way clustering: an unsupervised approach for gene expression data analysis. In: Proceedings of the 2nd IEEE international symposium on bioinformatics and bioengineering, pp 41–48

[32] Zeimpekis D, Gallopoulos E (2005) TMG: a MATLAB toolbox for generating term-document matrices from text collections. In: Kogan J, Nicholas C, Teboulle M (eds) Grouping multidimensional data: recent advances in clustering. Springer, Berlin, pp 187–210