Highly Relevant Routing Recommendation Systems for Handling Few Data Using MDL Principle

ABSTRACT

Many classification algorithms existing today suffer in handling many so-called "few data" instances. In this paper we show how to score query relevance with handle on few data by adopting a Minimum Description Length (MDL) principle. The main outcome is a strongly relevant routing recommendation system model (average(F)=0.72, M≥0.71; all scales of 1) supported by MDL based classification which is very good in handling few data by a large percentage margin of data degeneration (upto 50% loss).

CCS CONCEPTS

• **Retrieval tasks and goals** → **Clustering and classification**; *Sentiment analysis*; Recommender systems; • **Evaluation of retrieval results** → Relevance assessment;

KEYWORDS

MDL classification, relevance scoring, boosting factor

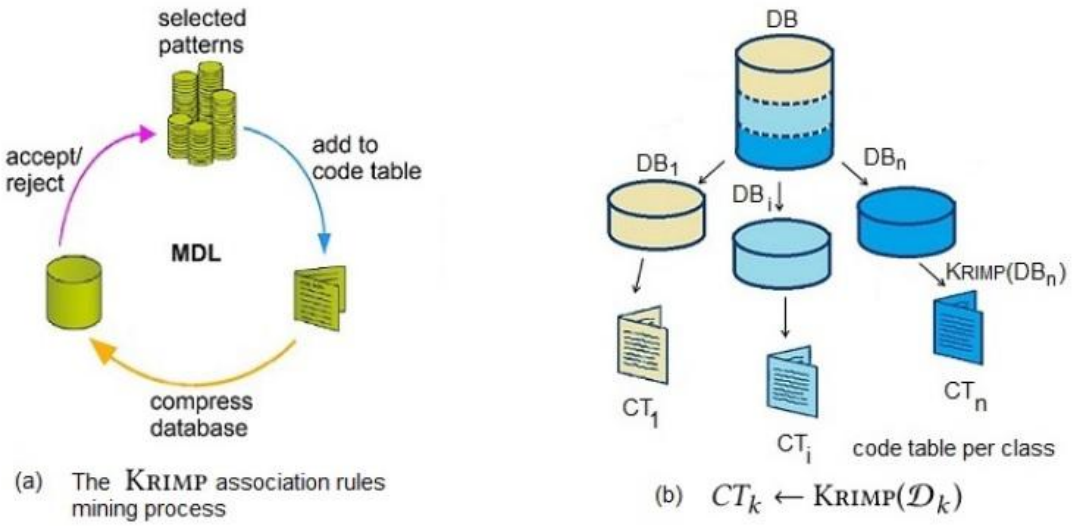
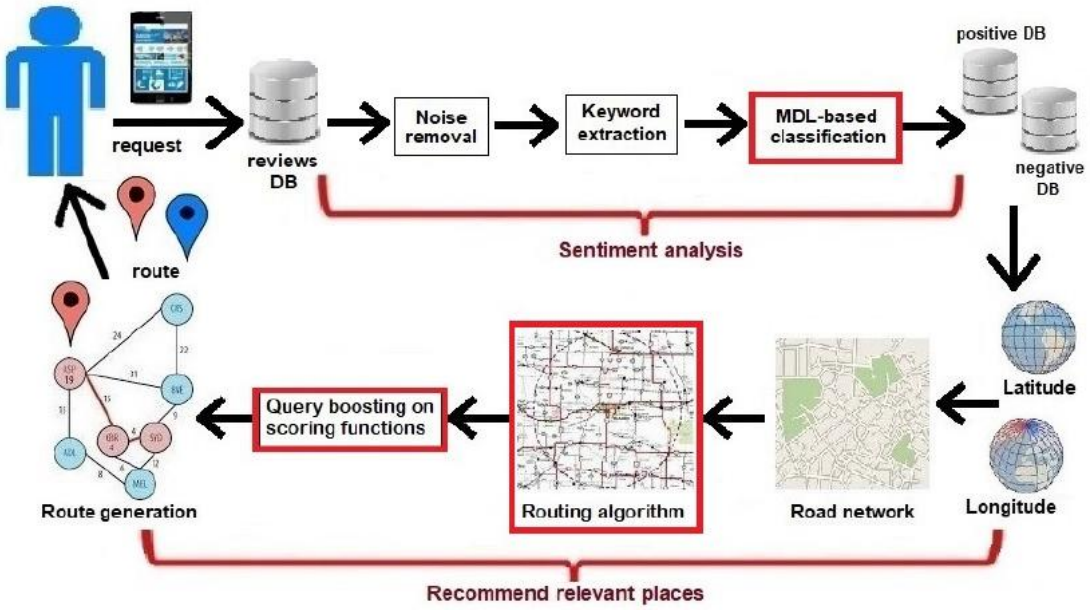


Figure 1: Preprocessing activities



5 CONCLUSIONS

In this work, we have focused on providing high quality recommendation system model using boosting factors in scoring function and the use of MDL principles on classification tasks. We have shown that using MDL-based classification we can handle few instances classification very well. Our findings indicate that improvements in classification can be achieved by using code table from large datasets. The performance of MDL-based classification degrades as the degeneration factor δ increases but it still outperforms other state of the art classification techniques. We permit upto 50% of data loss. Next, we have found that a proper selection of shortest routing algorithm (e.g. A* or Yen's) can provide further gains in quality. Finally, adding boosting factors can provide further gains in quality of the recommendation system model. The recommendation system model described in this paper can be implemented in various recommendation projects that use routing data and facing the few data instances problem on its knowledge source.

REFERENCES

[1] Judit Bar-Ilan, Mazlita Mat-Hassan, and Mark Levene. 2006. Methods for Comparing Rankings of Search Engine Results. *Comput. Netw.* 50, 10 (July 2006), 1448–1463. https://doi.org/10.1016/j.comnet.2005.10.020

[2] Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods* (1 ed.). Cambridge University Press.

[3] E. W. Dijkstra. 1959. A Note on Two Problems in Connexion with Graphs. *Numer. Math.* 1, 1 (Dec. 1959), 269–271. https://doi.org/10.1007/BF01386390

[4] Richard O. Duda and Peter E. Hart. 1973. *Pattern Classification and Scene Analysis.* Wiley.

[5] Ronald Fagin, Ravi Kumar, and D. Sivakumar. 2003. Comparing Top K Lists. In *Proc. of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '03).* Soc. for Industrial and Applied Mathematics, Philadelphia, PA, USA, 28–36.

[6] Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: The Definitive Guide.* O'Reilly.

[7] Guilherme Vale Menezes, Jussara M. Almeida, Fabiano Belém, Marcos André Gonçalves, Anísio Lacerda, Edleno Silva de Moura, Gisele L. Pappa, Adriano Veloso, and Nivio Ziviani. 2010. *Demand-Driven Tag Recommendation*. Springer Berlin Heidelberg, Berlin, Heidelberg, 402–417. https://doi.org/10.1007/978-3-642-15883-4_26

[8] Judea Pearl. 1984. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[9] J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[10] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic Keyword Extraction from Individual Documents. In *Text Mining, Applic. and Theory*. Michael W. Berry and Jacob Kogan (Eds.). John Wiley and Sons, Ltd, 1–20. https://doi.org/10.1002/9780470689646.ch1

[11] Börkur Sigurbjörnsson and Roelof van Zwol. 2008. Flickr Tag Recommendation Based on Collective Knowledge. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*. ACM, New York, NY, USA, 327–336. https://doi.org/10.1145/1367497.1367542

[12] Craig Silverstein, Hannes Marais, Monika Henzinger, and Michael Moricz. 1999. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum* 33, 1 (Sept. 1999), 6–12. https://doi.org/10.1145/331403.331405

[13] Matthijs van Leeuwen and Diyah Puspitaningrum. 2012. *Improving Tag Recommendation Using Few Associations*. Springer Berlin Heidelberg, Berlin, Heidelberg.

[14] Matthijs van Leeuwen, Jilles Vreeken, and Arno Siebes. 2006. *Compression Picks Item Sets That Matter*. Springer Berlin Heidelberg, Berlin, Heidelberg, 585–592. https://doi.org/10.1007/11871637_59

[15] Jilles Vreeken, Matthijs van Leeuwen, and Arno Siebes. 2007. Characterising the Difference. In *Proc. the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD '07)*. ACM, New York, NY, USA, 765–774. https://doi.org/10.1145/1281192.1281274

[16] J.Y. Yen. 1971. Finding the k shortest loopless paths in a network. *management Science* (1971), 712–716.