A Technical Report on the Evaluation of Text mining

Abstract

AI has received wide attention since AlphaGo defeated the Go world champion. Text mining, as an application of AI technology, has been highly anticipated because of its great potential. To provide an objective primer on text mining, this report presents a literature review, describes its general and specific operating principles and evaluates the technology in terms of effectiveness and efficiency. It was found that, currently, text mining is a mix of 7 domains with a general workflow that contains 5 steps. Furthermore, text mining is neither effective nor efficient. That is because, firstly, it is still an immature technology with a severe technical problem, nature language ambiguity, need to be conquered, and secondly, it may cause knowledge overload.

Contents

1. Introduction

Artificial intelligence (AI) is a technology that simulates human cognitive abilities through machines [1]. In 2016, a program called AlphaGo which is developed by Google defeated the former Go world champion by a final score of 4 to 1; A year later, an improved version of it defeated the world's top-ranked Go player at the time [1]. The application potential of AI has not only been fully proven but has also received widespread attention. Text mining technology, as an application of AI, can solve the problem of information overload. Talib et al. [2] define text mining as “a process to extract interesting and significant patterns to explore knowledge from textual data sources.”

The major issue of text mining is the complexi 2.1 of natural language, which exerts a negative influence on its effectiveness [2, 9, 11]. Ambiguity is one nature of natural language because it brings flexibility and usability to languages, so ambiguity cannot be eliminated entirely [11]. However, since ambiguity means that information which languages convey could be interpreted in various ways, for example, one word may have different meanings and different words may have the same meaning, it is very difficult for computers to understand natural language compared to humans [2, 11]. Although there have been numerous studies that are trying to address this problem, disambiguation techniques are still immature.

Even though there are still important technical challenges in text mining that have not yet been overcome, the potential of this technology is enormous, because it can bring a significant increase in efficiency. In the academic and research field, text mining is used for study identification in systematic reviews and extracting information from research articles [3, 4]; In business, it is possible to forecast market and intraday stock price trend, detect online review spam and uncover user satisfaction from online reviews [5]-[8]; When text mining is applied to social media, it could be used to discover different groups’ opinion on the same topic and help to identify cybercrime, such as cyberbullying, happening on the internet [9, 10]. Therefore, effectiveness and efficiency are currently the two main criteria for judging text mining.

The objective of this report is to describe how text mining works from whole to part, evaluate it using two criteria, effectiveness and efficiency, and give recommendations based on the evaluation.

2. Operating Principles

2.1. General principles

On the whole, techniques of text mining can be divided into seven groups: information extraction, information retrieval, document classification, document clustering, natural language processing, concept extraction, and web mining [2].

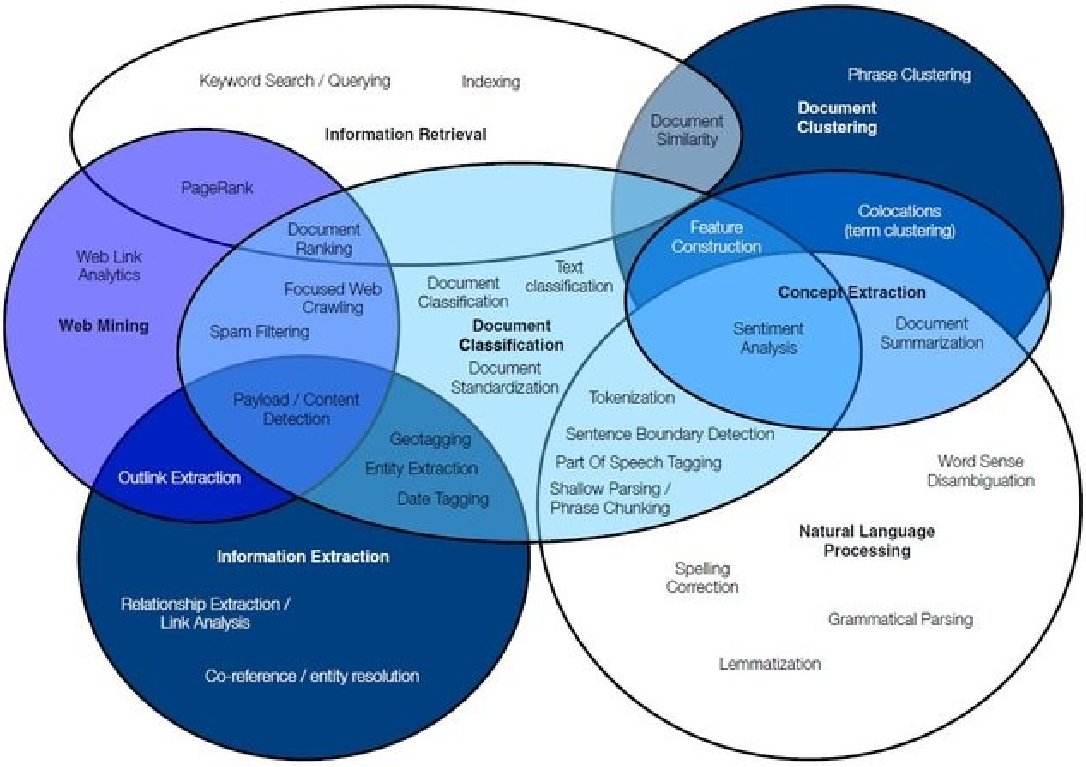

Fig. 1. Venn diagram for text mining techniques and their core functionalities [12]

Figure 1 associates different text mining techniques via their core functionalities, which presents a brief interrelationship within text mining.

2.2. Specific principles

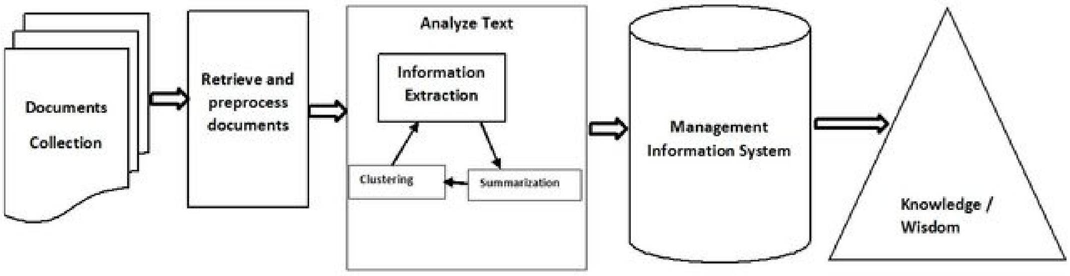

Fig. 2. Text mining process [13]

Figure 2 illustrates a common process of text mining. Overall, there are five stages in this process. To begin, texts are collected, and most of them are unstructured, which is hard for machines to analyse. Subsequently, texts have to be pre-processed before applying analytical techniques to them. In specific, depending on different text mining purposes, different characteristics of texts are needed, so other useless information should be cleansed. Afterwards, structured texts are analysed via a combination of techniques. The third stage shown in Figure 2 is a generic combination of techniques. Eventually, necessary information produced by the third stage is stored in the management information system, and by using pattern analysis, potential knowledge which users search for will come out.

3. Critical Evaluation

3.1. Evaluation for effectiveness

Effectiveness is used to judge whether desired outcomes can be produced. Expectations for outcomes may vary; the recognised one is the simplification of the text [1]-[4], [7]-[9].

At present, text mining is not an effective technology due to several issues it encountered. Firstly, the ambiguity of nature language is abstruse to machines [2, 9, 11]. For example, the meaning of synonyms, polysemy, and antonyms can only be determined by context, and abbreviations give changed meaning in different situations [2, 9, 11]. However, the ambiguity cannot be eliminated in advance, because of the flexibility and usability it brings to language; therefore, it is hard to write programs that can analyse raw text directly [11]. Secondly, one algorithm for text mining may not suitable for multiple languages [2, 9]. So far, although there are only a few tools that support multiple languages, many unsupported documents with important information are ignored [2, 9]. Thirdly, terminology for specific areas needs to be maintained by professionals [2]. Even though text mining is still immature, the foreseeable application potential can still prove its effectiveness in the future.

3.2. Evaluation for efficiency

Efficiency can be seen as the ratio of desired outcomes to costs. Different considerations of desired outcomes and costs can lead to different perceptions of efficiency of the same technology. For text mining, the recognised cost is time [1, 2, 4, 9, 11].

Currently, text mining is not an efficient technology. At first sight, text mining is an efficient technology, because it saves people time in reading and understanding text [1, 2, 4, 8, 9, 11]. Nevertheless, the knowledge produced by text mining is not necessarily valid. The reasons for this have been described in the previous section, so humans have to check the validity of the results, which takes time. Furthermore, assuming that text mining becomes an effective technology, then the number of results being produced may exceed the scope which humans can manage. The consequence will be that humans will have to spend more time managing these results. Therefore, if taking time spent on validation and management into account, text mining is not efficient.

3.3. Recommendations

Based on the above two-sided evaluation, here two recommendations are given. The first is about effectiveness. Some researchers avoided solving the issue of natural language complexity; they proposed that the document can be split into two parts, one for the original text and one for the structured text for ease of text mining [14, 15]. Although producing structured text will raise labour costs, it is a simple and effective solution to put text mining into practice as quickly as possible before overcoming the technical challenge of natural language ambiguity. Next on efficiency. Efficiency is effect-based, so the first step should be to improve the unsatisfactory effectiveness. After that, knowledge identification and management methods need to be developed in advance of the widespread use of text mining in order to prevent the knowledge overload caused by this technology.

4. Conclusion

4.1. Summary

In review, this technical report has described the operating principles of text mining and evaluated its effectiveness and efficiency. It was found that text mining uses 7 techniques as its technical foundation, and it has a general working procedure that overall has 5 steps. Furthermore, text mining is neither effective nor efficient since it is still an immature technology with technical problems that need to be conquered, and the knowledge overload it may cause.

4.2. Limitations

This report only draws on a small part of the literature on text mining, so it does not provide a detailed list of the unsolved technical challenges, nor does it provide a comprehensive evaluation of the technology on both advantages and flaws.

4.3. Recommendations

This is not an exhaustive technical report. Although contents are clearly structured, it is short and uninformative. However, it is for this reason that this report is well qualified for an initial introduction for groups of people who have no prior knowledge of text mining, such as students and companies. They can quickly grasp some concepts which are useful for them from the report. Hence, this report is worth reading for newcomers to text mining.

(1278 words)

References

[1] X. Tang and Y. Cheng, *Fundamentals of Artificial Intelligence*. Shanghai: ECNUP, 2018, ch. 1, pp. 9-14, ch. 7, pp. 121, 133, 134.

[2] R. Talib, M. K. Hanif, S. Ayesha, and F. Fatima, “Text Mining: Techniques, Applications and Issues,” *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 11, pp. 414-418, 2016. [Online]. doi: 10.14569/IJACSA.2016.071153 [Accessed Oct. 1, 2020].

[3] A. O’Mara-Eves, J. Thomas, J. McNaught, M. Miwa and S. Ananiado, “Using text mining for study identification in systematic reviews: a systematic review of current approaches,” *Systematic Reviews*, vol. 4, no. 5, 2015. [Online]. doi: 10.1186/2046-4053-4-5 [Accessed Oct. 5, 2020].

[4] S. A. Salloum, M. Al-Emran, A. A. Monem and K. Shaalan, “Using Text mining techniques for Extracting Information from Research Articles,” in *Intelligent Natural Language Processing: Trends and Applications*, K. Shaalan, A. E. Hassanien, and F. Tolba, Ed. Springer, 2018, pp. 373-397. [ebook]. doi: 10.1007/978-3-319-67056-0_18 [Accessed Oct. 1, 2020].

[5] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, “Text mining for market prediction: A systematic review,” *Expert Systems with Applications*, vol. 41, no. 16, pp. 7653-7670, 2014. [Online]. doi: 10.1016/j.eswa.2014.06.009 [Accessed Oct. 1, 2020].

[6] M.-A. Mittermayer, “Forecasting Intraday Stock Price Trends with Text Mining Techniques,” presented at the 37th Annual Hawaii International Conference. [Online]. Available: https://ieeexplore.ieee.org/document/1265201 [Accessed Oct. 1, 2020].

[7] R. Y. K. Lau, S. Y. Liao, R. C.-W. Kwok, K. Xu, Y. Xia, and Y. Li, “Text Mining and Probabilistic Language Modeling for Online Review Spam Detection,” *ACM Transactions on Management Information Systems*, vol. 2, no. 4, Dec. 2011. [Online]. doi: 10.1145/2070710.2070716 [Accessed Oct. 1, 2020].

[8] X. Xu and Y. Li, “The antecedents of customer satisfaction and dissatisfaction toward various types of hotels: A text mining approach,” *International Journal of Hospitality Management*, vol. 55, pp. 57-69, May 2016. [Online]. doi: 10.1016/j.ijhm.2016.03.003 [Accessed Oct. 3, 2020].

[9] S. S. Tandel, A. Jamadar, and S. Dudugu, “A Survey on Text Mining Techniques,” presented at International Conference on Advanced Computing &

Communication Systems. [Online]. doi: 10.1109/ICACCS.2019.8728547 [Accessed Oct. 1, 2020].

[10] A. Edwards, L. Edward, and A. Leatherman, “Text Mining and CyberCrime,” in *Text Mining: Applications and Theory*, M. W. Berry and J. Kogan Ed. Wiley, 2010, ch. 8, pp. 149-164. [Online]. doi: 10.1002/9780470689646.ch8 [Accessed Oct. 2, 2020].

[11] S. V. Gaikwad, A. Chaugule, and P. Patil, “Text Mining Methods and Techniques,” *International Journal of Computer Applications*, vol. 85, no. 17, pp. 42-45, Jan. 2014. [Online]. doi: 10.5120/14937-3507 [Accessed Oct. 1, 2020].

[12] W. He, “Examining students’ online interaction in a live video streaming environment using data mining and text mining,” *Computers in Human Behavior*, vol. 29, no. 1, pp. 90-102, Jan. 2013. [Online]. doi: 10.1016/j.chb.2012.07.020 [Accessed Oct. 1, 2020].

[13] S. Liao, P. Chu, and P. Hsiao, “Data mining techniques and applications – a decade review from 2000 to 2011,” *Expert Systems with Applications*, vol. 39, no. 12, pp. 11303-11311, Sep. 2012. [Online]. doi: 10.1016/j.eswa.2012.02.063 [Accessed Oct. 1, 2020].

[14] M. Gerstein, M. Seringhaus, and S. Fields, “Structured digital abstract makes text mining easy,” *Nature*, vol. 447, pp. 142, May 2007. [Online]. Available: https://www.nature.com/articles/447142a [Accessed Oct. 1, 2020].

[15] G. Mani and T. Hope, “Viral Science: Masks, Speed Bumps, and Guard Rails,” *Patterns*, vol. 1, no. 6, Sep. 2020. [Online]. doi: 10.1016/j.patter.2020.100101 [Accessed Sep. 20, 2020].

Index of comments

2.1	wrong citation order
5.1	how?
5.2	need examples to support
5.3	citation needed