Generating Competitive Intelligence Digests with a LDA-Based Method: A Case of BT Intellact

Qiang Wei(✉), Jiaqi Wang, Guoqing Chen, and Xunhua Guo

School of Economics and Management, Tsinghua University, Beijing 100084, China
{weiq,wangjq3.10,chengq,guoxh}@sem.tsinghua.edu.cn

Abstract. Internet has transformed the ways that organizations gather, produce and transmit competitive intelligence (CI), especially in the age of big data. This paper introduces a competitive intelligence digest generation method based on LDA topic modelling and representative text extraction. With the incorporated metric of perplexity, the proposed method is capable of automatic grouping of the texts and generating CI digests in an appropriate number of topics. Moreover, the method is applied to the context of BT Plc in the form of a case study, demonstrating its effectiveness in practical use.

Keywords: Competitive intelligence · LDA-based · Topic generation · Representative documents extraction

1 Introduction

Competitive intelligence (CI) such as market environmental dynamics, rivals' updates, techniques' hot spots, etc., plays a critical role in supporting executives and managers to make strategic decisions for an organization [1]. Nowadays, the Internet, as an information-rich open-source platform and an inter-organizational communications tool, has transformed the ways that organizations gather, produce and transmit competitive intelligence.

In the age of big data, competitive intelligence is generally hidden and should be discovered from various information sources online including news, business reports, surveys, financial reviews, etc., whereas traditional search tools and information retrieval methods can hardly provide satisfactory outcomes for competitive intelligence in an automatic and effective fashion. For instance, every day, a market researcher could easily collect/crawl a huge amount of rivals' data and market surveys, but is usually facing a problem of information overload due to the fact that the data/information is often sparse, conflicting, diverse or redundant, which makes CI difficult to generate and comprehend. In this regard, providing insightful digests (small and manageable sets of extracted important results/entries) of competitive intelligence (hereafter also referred to as CI digests) is considered meaningful and important for market researchers and then managers. Unlike traditional techniques such as information retrieval/summarization with manual manipulation and hand crafting, our work focuses on a LDA-based method for generating CI digests with an application in the context of British Telecommunications Plc (BT).

F.F.-H. Nah and C.-H. Tan (Eds.): HCIBGO 2016, Part I, LNCS 9751, pp. 341–349, 2016.
DOI: 10.1007/978-3-319-39396-4_31

BT is Britain's largest telecommunication company, whose business covers more than 170 countries and extensive entities. In meeting rapid market changes where new technologies and competitors are emerging, BT created an internal business intelligence unit called BT Intellact Department (BTID), aimed to collate and refine industry-related text information on the Internet for all employees in BT Group, and to continue daily updates of the information. The original sources of information BTID handles include online business news, trade journals/magazines, non-public information BT purchases from press operators and industry institutes, as well as internal research reports. Finally, CI digests are summarized under various topics and provided to end-users every week after users subscribe their preferred channels (i.e., labelled with topic tags). Though BTID has brought BT a solid competitive advantage through CI digests, handcrafting of human experts was heavily involved in the text analysis. The workflow for the service is illustrated in Fig. 1.

Apparently, BTID's service was practically valuable but encountered two challenges: (1) low efficiency of hand-crafting in the timely updating big-data environment; and (2) high hand-crafting burden for the BTID experts on not only clustering the huge amount of texts, but also extracting diverse and representative texts. This motivated us to develop a data-driven intelligent method.

In consideration of large-scaled and unlabeled data sources as well as their rapid updates, unsupervised clustering is deemed methodologically appropriate in processing and generating CI digests. Concretely, the Latent Dirichlet Allocation based (i.e., LDA-based) method is adopted in forms of text topic modelling, so as to effectively extract the valuable latent topics intelligently and group similar texts automatically, thus largely reducing the manual involvement [2].

The paper is organized as follows. Section 2 briefly overviews the related literature on competitive intelligence analysis and the LDA methods. Section 3 presents a LDA-based CI digest generation method. Section 4 analyzes a case on BT Intellact with the proposed CI digest generation. Finally, Sect. 5 provides concluding remarks and future work.

2 Literature Overview

Competitive Intelligence (CI) digests can be applied to various business areas. CI digests in marketing are to understand the latest market needs and users' feedbacks. Production

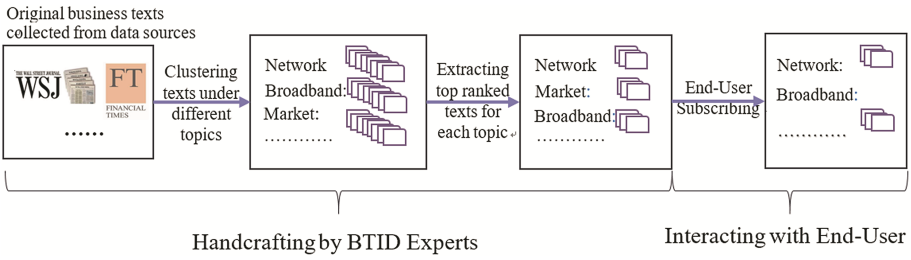

Fig. 1. Current workflow in BTID

departments can pick better suppliers and be informed of newest technologies of competitors with help of CI digests. Strategic CI digests can help top managers capture business insights to support their strategic decisions. Through concise and valuable information brought by CI digests, the overall efficiency of organizations could be greatly enhanced [1].

In CI digest generation, prior research efforts have resulted in a series of findings and techniques with regard to semantic modelling with natural language, representative information extraction and evaluation, competitive keyword suggestion and marketing method, semantic transitivity analysis as well as the corresponding information retrieval methods, text mining methods, etc.

Formally, given a set C of online collected n business texts, CI digest generation is to extract a small set of m texts, denoted as D, where $m << n$. Since the collected business texts are generally without explicit and structural labels, thus unsupervised clustering could be conducted, which can divide texts based on their similarities into several categories [3–6]. Furthermore, considering the semantic nature in CI digest generation, the texts can be grouped more effectively based upon their topic similarities rather than word similarities. Thus, in this spirit, the well-known LDA methodology is regarded suitable [4]. LDA is a three-level Bayesian clustering for latent topic modeling [7–11].

For grouping unstructured and latent text topics, LDA possesses the following merits [12–17]. First, LDA's effectiveness for large-scale text clustering is very desirable, and its efficiency performance is also acceptable. Second, its probability model is solid, showing strong adaptability and scalability in many applications. Third, it is conscious of the influence of the text structure on text meaning in addition to word frequency, which could dig out the hidden semantics of texts. Fourth, it allows for characteristics of multi-topics of texts, which conforms to practical cases.

3 A LDA-Based Intelligent CI Digest Generation Method

Generally, given a set C of n texts (business news, reports, blogs, surveys, etc.) collected from open sources, a LDA-based semantic text mining method is used to extract a small set D of m texts, where $D \subseteq C$, and the text in D is the most representative text with respect to its corresponding category in C.

Concretely, the CI digest generation process is composed of two stages, i.e., LDA-based clustering with topic tag assignment and representative texts extraction for each clustered category.

In the LDA-based clustering stage, first, each text in C is preprocessed and parsed, represented by a vector of extracted keywords along with their frequencies, based on which the corresponding LDA semantic model can be built. Second, all keywords represented by a vector of keywords with latent semantic relevance will be clustered into different categories based on the LDA model, i.e., m clusters/categories are generated. Third, for each cluster with multiple keywords along with latent semantic relevance, an appropriate topic ID or tag will be assigned to each category. Thereafter, the texts are automatically clustered into m categories. Obviously, its efficiency outperforms manual labelling operation.

Next, in the representative texts extraction stage, the text with highest LDA relevance in each category (e.g., containing usually tens or hundreds of texts) could be extracted as the most representative text. Thus, the whole set of a CI digest is generated with m texts with respect to the original set of n texts.

The general framework of the proposed method is shown in Fig. 2.

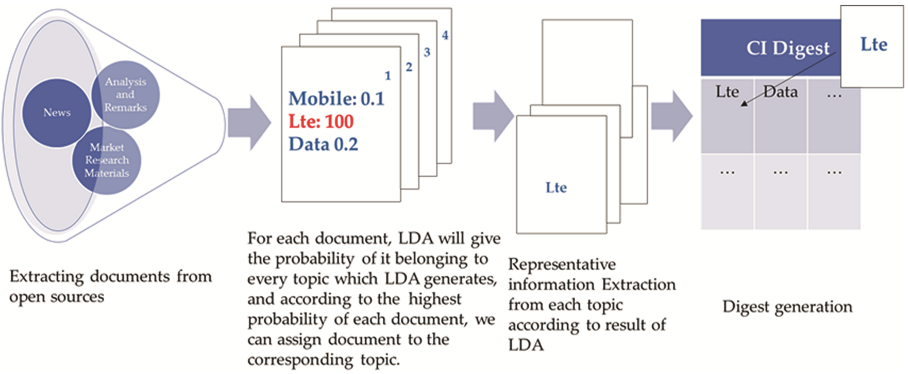

Fig. 2. General framework of the LDA-based CI digest generation

During the process of the proposed method, the number of topics, i.e., m, is to be predetermined. If m is set too small, i.e., too few topics, the derived CI digest will be less informative; on the contrary, too big m means that too much tedious information will be retained in the derived CI digest. Therefore, to determine an appropriate m value significantly affects the final results. However, due to the users'/experts' cognitive difficulty in getting a whole picture for totally n texts, it is hard for them to configure an appropriate value of m. Therefore, according to Blei et al. [4], a metric, i.e., *Perplexity*, could be used here for helping determine the m value by assessing the quality of topic model through its prediction effect. *Perplexity* is as defined in Eq. (1). The lower *Perplexity* is, the more representative this topic model is.

$$Perplexity(D) = \exp\left\{\frac{-\left(\sum_{d=1}^{m} \log(p(w_d))\right)}{\sum_{d=1}^{m} N_d}\right\}, \tag{1}$$

where d is a derived topic, N_d is the number of words in d, and $p(w_d)$ is the probability of every word in d, m is the number of topics. Thus, by minimizing the *Perplexity* of the original set of D, the appropriate number of topics, i.e., m, can be derived.

By integrating the *Perplexity* optimization process into stage 1, the method is finally devised. With this proposed method, a CI digest could be automatically and intelligently extracted from a large amount of original texts. For the example of BTID workflow in Fig. 1, if the proposed method could be integrated, the workflow could be improved as shown in Fig. 3.

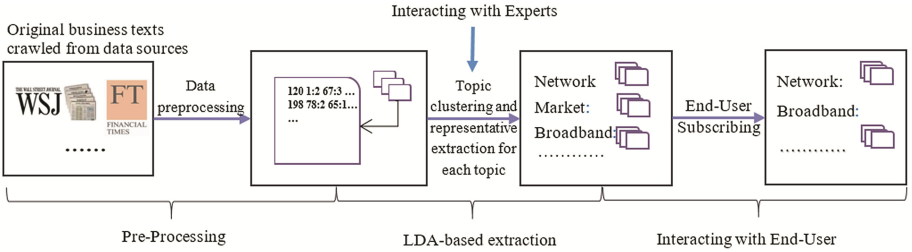

Fig. 3. Improved workflow in BTID

Figure 3 shows that, first, the crawled texts could be pre-processed into structured data, which can be used in the LDA-based extraction. Then, with the LDA-based extraction, the number of topics, and all the topics as well as representative texts for corresponding categories could be derived. Theoretically, this step can be conducted automatically without human intervention. Nevertheless, in our method, an interface is designed to interact with experts (e.g., BTID experts) for investigating the results and necessarily adjusting, e.g., tag names, text assignments, according to their domain knowledge, which provides more flexibility and robustness of the workflow.

It should be emphasized that human intervention integrated in the method does not weaken the contribution of the method. First, exogenous knowledge is only used to name the tags, which does not affect the automation of the method. Second, as a typical decision support process, the LDA-based extraction does not substitute experts' knowledge, but significantly augment experts' insights on the crawled texts, essentially leading to better CI digest generation.

Finally, with the experts-improved results, the end-users can browse related CI digest by subscribing preferred topics. To further demonstrate the effectiveness of the proposed method and the improved workflow, one analytic case of BT is discussed in the next section.

4 A Case of BTID CI Digest Generation

BTID produced weekly Competitive Intelligence digests to push to the employees within the corporation according to their subscription since BTID was established, but the whole process including topic generation, text clustering and representative text extraction were all processed by hand or simple tools. Therefore, it was then considered meaningful to improve the workflow with the proposed method that is of business analytics nature. This section introduces the analysis on the real data from BTID.

In a joint research project with BT, more than 1,300 full business texts were provided by BTID, covering the topics such as the analysis of competitors, telecom & IT industry, new techs, market environment, government policies, sales, etc., handcrafted by BTID experts. On average, there are 560 words per text. After filtering with explicitly noisy texts, 1,277 full texts were used for analysis.

Subsequently, data preprocessing was firstly conducted, i.e., stop words deletion, case changing, title weighting (i.e., 3 times weighting on title was used, which is a typical

configuration for related text mining [18]), input format preparation (i.e., each text was transformed as the number of keywords as well as a list of pairs of keyword and frequency), finally a vocabulary for corpus about the texts was constructed.

Furthermore, the widely-used JibbsLDA package was used for conducting the LDA modeling, with the typical configuration, i.e., alpha = 0.01, beta = 0.1, etc. Before retrieving the LDA semantic models, the appropriate number of topics (i.e., m) had to be determined by minimizing *Perplexity*. Different m (m = 1, 2, …, 20) values were tested, with results as shown in Fig. 4.

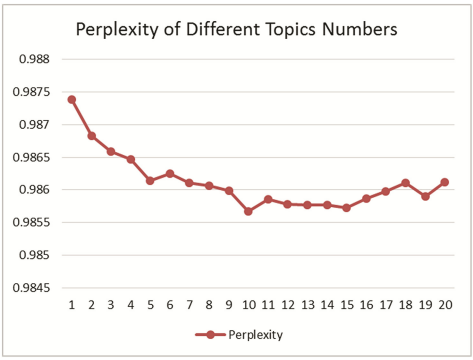

Fig. 4. Perplexity of different topic numbers

Figure 4 exhibits that the best topic number of the case was around 10. Since the *Perplexity* value decreased continuously before $m = 10$ and was stable and higher after $m = 10$, in the following discussion of the case, $m = 10$ was used.

Moreover, based on the 10 topic configuration, to be more understandable, the number of top keywords for each topic was set as 5 for LDA modeling and topic generation. The topics and corresponding keywords from LDA modeling are represented in Table 1. It should be noted that, LDA modeling itself can only present the topic IDs not topic names. In this case, experts were involved to help generalize an appropriate topic tag for each category. In addition, experts were also asked to help check whether obtained topics, extracted keywords, and categorized texts below each topic were reasonable based on their industry knowledge, and they were authorized to make necessary adjustments accordingly. Here, with LDA topic modeling, the experts only endowed topic tags without other intervention in particular, saving the vast amounts of efforts in otherwise human-involved text preparation, reading, and grouping, which nowadays becomes more and more impossible when huge volume of data pertains in practice.

Table 1. Generated topics from BT corpus

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Topic category with 5 keywords	Airline	India	Data	Ford	Bank
	United	Crore	Services	Fraud	Capital
	Airport	Patients	Business	Blackberry	Tests
	Flight	Health	Cloud	Aluminum	Financial
	Aircraft	Delhi	Technology	Vodafone	Lloyds
Topic tag	Airline	Indian Health	Data Service	Partner	Finance
	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Topic category with 5 keywords	Company	Mobile	Government	People	Car
	Market	Broadband	Public	Social	Vehicles
	Year	BT	People	Online	BMW
	Billion	UK	Law	Media	Engine
	Sales	EE	Police	Facebook	Power
Topic tag	Market	Mobile	Government Policy	Social Media	Vehicles

Compared with previous hand-crafted topics, the LDA-based method could generate most of the topics listed by experts in BT, such as Market, Government Policy and Social Media, Mobiles, Data Services, Vehicles, etc. In addition to these existing ones, some new topics appeared such as Airline, Finance, Indian Health and so on. After discussions with telecom experts, these new topics (though they were not listed by experts with handcrafting) were also acknowledged as BT's focuses at that time, reflecting the power of the proposed method for finding more novel and useful topics.

Table 2. Representative texts' index for every topic in BT corpus

Topic	1	2	3	4	5	6	7	8	9	10
Index of text	1247	851	6	876	689	417	1276	802	242	391

With Table 1, the representative texts for the 10 topics were further extracted respectively. Due to the limitation of space, Table 2 only lists the index of the 10 texts, which form the final derived CI digest for this case. Moreover, for illustrative purposes, the texts, i.e., No. 802 and No. 391, for topics "Government Policy" and "Vehicle", are listed in Table 3.

Finally, to justify the quality of the final derived results, a TREC test was conducted, where each derived text was investigated by 3 human experts to assess whether the content was consistent with its assigned topic tag. As a result, an over 90 % accuracy was reported in the test, further showing the effectiveness of the proposed CI digest generation method.

Table 3. Partial CI digest of BT

CI digest of BT	
Government Policy	Vehicles
Warning Over Planning Policy: The Government's flagship planning policy is leading to "inappropriate and unwanted housing development", MPs have warned. The cross-party Communities and Local Government Committee also raised concerns that town centres were not being given proper protection against the threat from large out-of-town retail developments. They called for the Government to scrap rules allowing small shops and offices to be converted to housing without the need for planning permission, arguing that the changes could lead to town centres becoming "an unattractive place to visit or, indeed, live"…	Germany: BMW 2 Series Coupe to feature new entry-level engines from March 2015. From March 2015, new entry-level engines, a further four-wheel drive model and additional equipment options will increase the diversity of features available for the BMW 2 Series Coupe. With the market launch of the new BMW 218i Coupe, a three-cylinder petrol engine from the BMW Group's latest engine family will be featured for the first time in the brand's sporty and elegant compact model…

5 Concluding Remarks

In this paper, a Competitive Intelligence (CI) digest generation method has been introduced to help organizations effectively and intelligently generate CI digests, significantly alleviating the burden of human work in text analytics and semantic modeling with huge volume of data. The proposed method is composed of two parts, namely, LDA-based topic modeling and representative texts extraction, where the metric of perplexity has been incorporated into the determination of the topic modeling quality. Moreover, a case of BTID CI digest generation has been illustrated and analyzed with the proposed method, showing the effectiveness of the proposed method.

Future work will be carried out in two respects. One is to apply the method to other large-scaled business environments; the other is to develop an incremental strategy for timely updating environments.

Acknowledgements. The work was partly supported by the National Natural Science Foundation of China (71490724/71110107027/71372044) and the Tsinghua-BT Advanced ICT Lab at Tsinghua University. The authors highly appreciate the support and cooperation of BT and Dr. Quan Li at BT China Research Centre for the work.

References

1. Teo, T.S., Choo, W.Y.: Assessing the impact of using the Internet for competitive intelligence. Inf. Manag. **39**(1), 67–83 (2001)
2. Zhe, G., Dong, L., Qi, L., et al.: An online hot topics detection approach using the improved ant colony text clustering algorithm. J. JCIT **2**, 243–252 (2012)

3. Sathiyakumari, K., Manimekalai, G., Preamsudha, V., et al.: A survey on various approaches in document clustering. Int. J. Comput. Technol. Appl. (IJCTA) **2**(5), 1534–1539 (2011)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
5. Sahoo, N., Callan, J., Krishnan, R., et al.: Incremental hierarchical clustering of text documents. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp. 357–366. ACM (2006)
6. Young, S., Arel, I., Karnowski, T.P., et al.: A fast and stable incremental clustering algorithm. In: 2010 Seventh International Conference on Information Technology: New Generations (ITNG), pp. 204–209. IEEE (2010)
7. Bradley, P.S., Fayyad, U.M., Reina, C.: Scaling clustering algorithms to large databases. In: KDD, pp. 9–15 (1998)
8. Farnstrom, F., Lewis, J., Elkan, C.: Scalability for clustering algorithms revisited. ACM SIGKDD Explor. Newsl. **2**(1), 51–57 (2000)
9. O'callaghan, L., Meyerson, A., Motwani, R., et al.: Streaming-data algorithms for high-quality clustering. In: ICDE, p. 0685. IEEE (2002)
10. Hulten, G., Spencer, L., Domingos, P.: Mining time-changing data streams. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 97–106. ACM (2001)
11. Kifer, D., Ben-David, S., Gehrke, J.: Detecting change in data streams. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases, VLDB Endowment, vol. 30, pp. 180–191 (2004)
12. Zhong, S.: Efficient streaming text clustering. Neural Netw. **18**(5), 790–798 (2005)
13. Banerjee, A., Basu, S.: Topic models over text streams: a study of batch and online unsupervised learning. In: SDM 7, pp. 437–442 (2007)
14. Maskeri, G., Sarkar, S., Heafield, K.: Mining business topics in source code using Latent Dirichlet Allocation. In: Proceedings of the 1st India Software Engineering Conference, pp. 113–120. ACM (2008)
15. Canini, K.R., Shi, L., Griffiths, T.L.: Online inference of topics with Latent Dirichlet Allocation. In: International Conference on Artificial Intelligence and Statistics, pp. 65–72 (2009)
16. Bíró, I., Siklósi, D., Szabó, J., et al.: Linked Latent Dirichlet Allocation in web spam filtering. In: Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web, pp. 37–40. ACM (2009)
17. Blei, D., Hoffman, M.: Online learning for Latent Dirichlet Allocation. In: Neural Information Processing Systems (2010)
18. Banerjee, S., Ramanathan, K., Gupta, A.: Clustering short texts using wikipedia. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 787–788. ACM (2007)