Customer Reviews Analysis Based on Information Extraction Approaches

Haiqing Zhang, Aicha Sekhari, Florendia Fourli-Kartsouni, Yacine Ouzrout, Abdelaziz Bouras

Customer Reviews Analysis Based on Information Extraction Approaches

Haiqing Zhang[1], Aicha Sekhari[1], Florendia Fourli-Kartsouni[2], , Yacine Ouzrout[1], and Abdelaziz Bouras[3]

1. DISP laboratory, University Lumière Lyon 2, France, 160 Bd de l'Université 69676 Bron Cedex
2. Hypercliq, Pradouna 57, Athina 11525, Greece
3. Computer Science Department - Qatar University, ictQATAR, Box. 2731, Doha, Qatar
{aicha.sekhari, yacine.ouzrout,Taha.Elhariri}@univ-lyon2.fr
abdelaziz.bouras@qu.edu.qa; haiqing.zhang.zhq@gmail.com; f.fourli@hypercliq.com

Abstract. The existing information extraction approaches are generally analyzed and then categorized into several groups based on the superiority and the intelligence of the approaches as well as their capability to solve complex problems. Two practical approaches are provided to clarify how to use the information extraction solutions to obtain the valuable information from numerous reviews. The first approach is to support the front-end services in the EASY-IMP project. The customer preference and the optimum interest of customers is determined based on TF-IDF approach. Roughly 100,000 pages have been analyzed and the customer preference is studied based on the most relevant keywords. However, TF-IDF approach limits on the capability to provide the personalized infromation, which can only obtain the restricted information based on weights calcualtion. In order to extract more efficient customerized infromation, an opinion mining algorithm is proposed. The proposed algorithm aims to obtain sufficient information extraction results and reduce the complexity and running time of information extraction by jointly discovering the main opinion mining elements. The analyzed reviews show that the proposed algorithm can effectively and simultaneously identify the main elements.

Keywords: Information Extraction; TF-IDF; Opinion Mining; Dependency Relations; Part-of-Speech

1 Introduction

EASY-IMP[1] project is founded to develop methodologies, tools, and platforms for design and production of personalized meta-products by combining wearable sensors embedded into garment based on mobile and web-based technologies. A meta-product means a customer driven customizable entity that integrates sensory and computing units, leading to a paradigm shift from mass production to intelligent, over-the-web configurable products. The widely used Web communication on mobile and web-

[1] http://www.easy-imp.eu/

based technologies in this project has dramatically changed the way individuals and communities express their opinions on meta-products. More and more reviews are posted online to describe customers' opinions on various types of products. These reviews are fundamental bits of information to support both firms and customers for making correct decisions. The features and attributes of a product extracted from online customer reviews can be used in recognizing the strengths and weaknesses of the heterogeneous products for firms. While customers do not always have the ability to wisely choose among a variety of products in the market, they commonly seek product information from online reviews before purchasing a new product. Moreover, the number of reviews grows rapidly, which becomes impractical if analyzing the reviews by hand. If features and the related opinions can be obtained from the massive reviews and then firms will gain great benefits by using the extracted information to evaluate how and where to improve the product through the product development process. Hence, in this paper, we have studied the information extraction approaches to analyze the reviews of meta-products.

The **i**nformation **e**xtraction (IE) task is to identify the entities, relations between objects, and obtain the relevant features of the identified entities. Based on (McCallum, 2005), The IE tasks are categorized in five groups in terms of segmentation, classification, association, normalization, and de-duplication. In order to extract the structured data from haphazard, noisy, and unstructured data to complete the IE tasks, the research works also adopt the previous techniques such as machine learning, data mining, information retrieval, and computer linguistics to solve the IE tasks.

However, fully addressing IE is a tough problem that the existing proposed algorithms can only solve a small part of IE tasks from the emergence of IE till now. In order to better comprehend the advantages and the capabilities of IE, we will give two practical applications by adopting IE solutions. This paper is structured in the following way: A brief study and categorize the existing IE solutions in section 2. Section 3 provides an application that comes from EASY-IMP project, which supplies end-user services based on the TF-IDF (Information Retrieval) approach. In order to more intelligently and more automatically extract the customer information from the reviews, section 4 proposed an opinion mining extraction algorithm to jointly extract features, opinions, and feature-opinion relations to reveal the strengths and weaknesses of the products' attributes; meanwhile, some extracted information is given based on the proposed algorithm. Section 5 concludes the work.

2 Literature Review

2.1 The Representative Approach of Information Retrieval: TF-IDF

The traditional techniques of IE mainly refer to information retrieval techniques, which are based on key word searches to figure out the most likely document or term by the searcher. In order to complete this task, the weights of the documents must be calculated to answer which one can best satisfy the searching query. The methods are used to assign the weights of terms are Binary Weights (Salton et al., 1983), Raw term frequency (Paltoglou and Thelwall, 2010), TF-IDF (Term Frequency-Inverse document frequency) (Hiemstra, 2000), etc. Particularly, TF-IDF is the most

commonly used method for web search tasks that orders the documents or terms based on the relevance to the searched query. Therefore, we will give a detailed explanation of TF-IDF.

Term frequency (tf_{xi}) is used to measure the term density in a document (D_x), which means the frequency of term T_i in document D_x. Inverse document frequency (IDF) is used to measure the discriminating ability of a term, which means the rarity of the term across the whole documents. Based on Aizawa (2003), the theoretical justification of TF-IDF shows that the optimal calculation of IDF for document retrieval is:

$$idf_{ti} = \log\left(\frac{n}{df_{ti}}\right) \tag{1}$$

Where, n is the total number of collected documents, df_{xi} is the total number of collected documents (D_t) that contain searched term T_i. And then, the formula that is used to express the term weights obtained by TF-IDF is shown as follows:

$$w_{xi} = tf_{xi} \times idf_{ti} = tf_{xi} \times \log\left(\frac{n}{df_{ti}}\right) \tag{2}$$

For instance, the term frequency tables for two documents are shown in Table 1. Then the calculation of tf-idf for two terms "Cricket" and "Grappling" is given in the following:

$$w_{\text{Cricket}}(\text{D1}) = tf_{xi} \times idf_{ti} = 2 \times \log\left(\frac{2}{2}\right) = 1 \times 0 = 0 \tag{3}$$

$$w_{\text{Grappling}}(\text{D2}) = tf_{xi} \times idf_{ti} = 4 \times \log\left(\frac{2}{1}\right) = 4 \times \log 2 \approx 1.2040 \tag{4}$$

Table 1. Example data: Term, Term Frequency and documents

Document 1		Document 2	
Term	Term frequency	Term	Term frequency
Cricket	2	Cricket	1
Rugby	1	Grappling	3

2.2 Morden Information Extraction Solutions

The modern information extraction solutions differ from the traditional techniques that extract the most important facts about features, entities, and relations from various documents (which may be combined by multiple languages). The obtained important facts are usually used to analyze the changing trend of reviewers' preference and recommendation, the summary of the document, and serve the new products development. The main modern information extraction solutions are categoried in the following:

- **Statistical approaches** (Dey and Verma, 2013; Blei et al., 2003;Blunsom, 2004): supervised or unsupervised to learn the properties or attributes of text;

classification of content into various categories through analysis of human-tagged labeled samples; extraction of hidden topics or grouping similar content. One of the representative methods is hidden Markov model (McCallum et al., 2000). This method calculates the probability that from one state to another based on the theory of probabilities, and hence obtaining the probabilities of several words emerging together.

- **Natural language processing** approaches, which mainly contain three main components are shown in the following: 1). Taggers: POS (part-of-speech)(Tsuruoka et al., 2005), which is used to understand the structure of the sentences. 2). Parses (McDonald et al., 2005; Nivre, 2005;De Marneffe et al., 2006): Analyze whole sentence structures and try to derive semantic relationships among the components of a single sentence. To identify finer grained emotions like wish, anger, fear etc. 3). Named Entity Recognizer (Nadeau and Sekine, 2007), which is a special category of NLP tools that employ pattern recognition techniques to extract named entities from documents. Named Entities include names of people, places, organizations, product models, time (money) -values, email addresses; telephone numbers, etc.
- **graph(or Tree)-based method** (Litvak and Last, 2008): The type approach mainly has three parts that include sentence structure analysis, constructing graph database, and graph similarity for merging.
- **regular expressions** (Li et al., 2008): Compile the regular expressions to explain the sentence pattern.
- **machine learning** (Aggarwal and Zhai, 2012): Define the specific rule or parameter to automatically extract the information.

3. The Proposed Front-End Services for EASY-IMP Project Based on TF-IDF Approach

A number of web-based services have been developed based on the EASY-IMP project. The users can build the profiles, discover the most suitable Meta-Products (MP) based on the related requirements, and then configure the MPs based on the customer preferences. In order to overcome the challenges related to the complexity and the creativity of the MPs, the front end services are provided to support customer profile building and personalized recommendation based on TF-IDF approach. The modules of the provided front-end services are briefly shown in Figure 1.

The front end services focus on developing the parts of ‘user profiling’ and ‘MP recommendation’. In order to more accurately reveal the user preference, the information including user interests and product preferences should be obtained. Once the basic user information has been retrieved, the user preferences are determined by the frequent items that have been predefined in a list of keywords with respect to a specific subject. The keywords in the social media are automatically generated through analysis of a large corpus of Facebook pages that related to the defined subjects, and the most relevant keywords are identified and selected as the preferred. Roughly 100,000 pages related to the topics of ‘Fitness’, ‘Exercise’, ‘Running’, and ‘Cycling’ have been obtained by using the Facebook Graph API. Text processing approaches have been used to do preprocessing. TF-IDF is adopted to calculate the relevance weights among words.

In short summary, the EASY-IMP front-end services have been created to provide the useful information about MPs development. The user profiling is built based on basic user information and the inferred user interests. The analyzed 100,000 Facebook pages has proved that the word weight determination base on TF-IDF approach can provide a list of meaningful keywords for accurately classifying new pages based on the predefined user preferences.

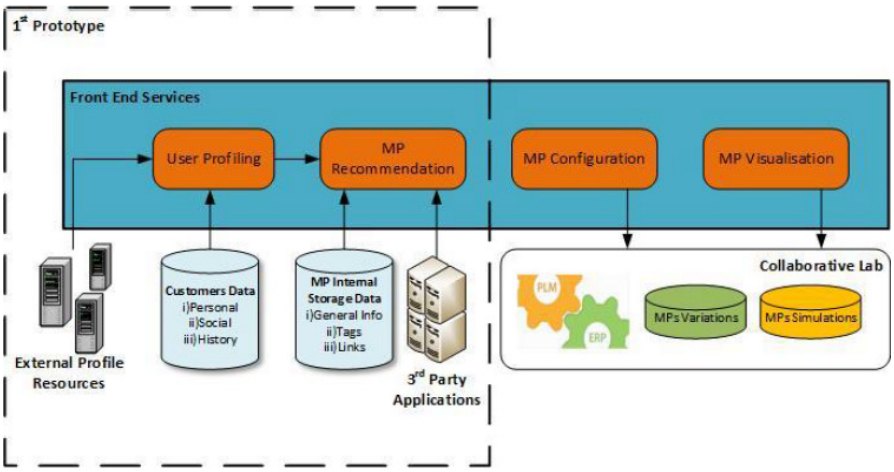

Fig. 1. The Modules of Front-End Services

3. The Proposed Opinion Mining Extraction Algorithm to Jointly Execute Opinion Mining Extraction Tasks

3.1. Information Extraction Sextuple

On the basis of keywords retrieval, in order to more intelligently discover the customers' preferences, more useful information will be obtained. The essential elements of customer reviews contain the features, with the opinion it expresses, and the relations between features and opinion expressions. The necessary information of reviews is defined as a quintuple in (Liu and Zhang, 2012), we extend the quintuple into a sextuple by adding the relations among features and opinions, which is shown as (e_i, f_{ij}, oo_{ijkl}, r_{ijkl} ,h_k, t_l), where e_i is the name of an entity; f_{ij} is a feature of e_i; oo_{ijkl} is the opinion expression on feature f_{ij} of entity e_i; r_{ijkl} is the sets of feature-opinion relation extraction, feature-feature relation extraction, and opinion-opinion relation extraction; h_k is the opinion holder; and t_l is the time when the opinion is expressed by h_k. This definition can provide a basis for transforming unstructured text to structured data in the following sections. The added attribute r_{ijkl} can be used to summarize the overall attitude of the whole review and reflect the opinions with respect to a specific feature.

3.2. Information Extraction Rules Defined Based on Dependency Relations

The extraction is mainly between features and opinion words. For convenience, some symbols are defined to be able to reuse them easily. The relations between opinions and features are defined as FO↔Rel, between opinion words themselves are

OO↔Rel, and between features are FF↔Rel. Six basic extraction tasks are defined to separate information extraction: (1). Extracting products' features by using opinion words (FO↔Rel); (2). Retrieving opinions by using the obtained features (OF↔Rel); (3). Extracting features by using the extracted features (FF-Rel); (4). Retrieving opinions based on the known opinion words (OO-Rel). (5). Extracting products' features by using both the extracted opinion words and the related features; (6). Extracting opinions based on the extracted opinions and features. The added two more tasks focus on implicit dependency relations especially for long distance dependency. Six catalogues of running rules are clarified for the proposed six tasks and the detail analysis is depicted in Table 2.

In Table 2, o (or f) represents for the obtained opinions (or product features). {O} (or {F}) is the set of known opinions (or features) either given or obtained. POS (O/F) means the POS information that contains linguistic category of words such as *noun* and *verb*.{NN, NNS, JJ, RB,VB} are POS tags to describe opinions or features. O-Dep represents the opinion word O depends on the second word based on O-dep relation, F-dep means the feature word F depends on the second word through F-dep relation. MR={nsubj, mod, prep, obj, conj,dep}, 'mod' contains {amod, advmod}, 'obj' contains {pobj, dobj}, which are dependency relations describing relations among words. Finally, the rules are formalized (we only show the main rules in this paper) and employed to extract features (f) or opinion words (O) based on the previously defined six tasks.

Table 2. Simplified Rules for Features and Opinion Expressions Extraction

Rule	Input	Representation Formula	Output	Example
R1	O	$O \xrightarrow{\text{Depend (O-Dep)}} F;$ where, $O \in \{O\}$, O-Dep $\in \{MR\}$, $POS(F) \in \{NN, NNS\}$	f=F; {FO}	Canon PowerShot SX510 takes **good** *photos*. (good→amod→photos) (Figure 2) The **images** are **excellent**. (excellent←nsubj←images)
R2	F	$O \xrightarrow{O-Dep} F;$ s.t. $F \in \{F\}, POS(O) \in \{JJ, RB, VB\}$	o=O {FO}	Same as $R1_1$, *photos* as the known word and *good* as the extracted word.
R3	F	$F_{i(j)} \xrightarrow{F_{i(j)}\text{-Dep}} F_{j(i)}$ s.t. $F_{j(i)} \in \{F\}, F_{i(j)}$-Dep $\in \{conj\}; POS(F_{i(j)}) \in \{NN, NNS\}$	f=F {FF}	It takes breathtaking **photos** and great **videos** too. (photos→conj→videos)
R4	O	$O_{i(j)} \xrightarrow{O_{i(j)}-Dep} O_{j(i)}$, s.t. $O_{j(i)} \in \{O\}$, $O_{i(j)} - Dep \in \begin{Bmatrix} advmod, \\ conj \end{Bmatrix}$, $POS(O_{i(j)}) \in \{RB\}$	o=O{OO}	Canon PowerShot *SX510* takes significantly **better** indoor *photos*. (better←advmod←significantly) This camera is **light** and easy to hold. (light←conj←easy)

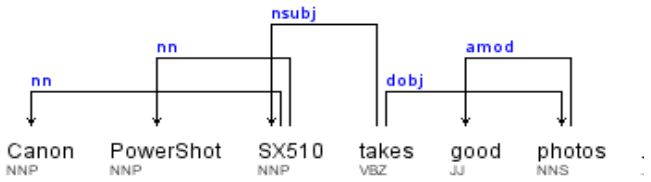

Fig. 2. The dependency structure for the sentence: Canon PowerShot SX510 takes good photos

3.3. Opinion Mining Extraction Algorithm

Table 3 shows the detailed opinion mining extraction algorithm. The initial values of the proposed algorithm are shown as: opinions dictionary O, the opinion degree intensifiers OD, and the review data RD. The opinion dictionary is based on Hu and Liu (2004) and the opinion degree intensifiers are defined by the authors. This algorithm adopts a single review from customers as the basic analysis unit. The products' features should be unique, while the opinion words to describe the features can be reused in each review. The algorithm stops when no more new features can be found.

Table 3. Algorithm: opinion mining extraction algorithm

Algorithm Opinion_Mining_Extraction()

Input: Opinion word dictionary O, Opinion Degree Intensifiers OD, Review Data: RD
Output: The set of features F, the set of expanded opinion words EO, the opinion polarity (or orientation) for a product: OW

BEGIN

1. Expanded opinion words: EO= ∅ ; F= ∅ ; ODI= ∅ ;
2. **For** each dependency parsed review RD_k
3. // Obtaining the initial opinion words and intensifier degree words in RD_k based on the dictionaries of O and OD
4. **for** each word tagged JJ,RB, and VB in RD_k
5. Traversing the RD_k, and extracting the opinion words (OP_i) if they are appearing in O; i++;
6. Extracting new opinion words {OP_j} in RDk by using the Rules R41-R42 based on extracted opinion words {OP_i}; j++;
7. Inputting the obtained OP_i and OP_j into EO, and then EO={OP[1,…,i] , OP[1,…,j] }(for short EO={ $OP_{1\text{-}i}$, $OP_{1\text{-}j}$ });
8. Traversing the RD_k, and extracting the degree intensifier words (DW_d) if they are appearing in OD;
9. Inputting the obtained DW_d into ODI, and then ODI={$DW_{1\text{-}d}$}; d++;
10. **End for**
11. //Extracting the features based on the obtained initial opinion words and opinion degree intensifier words
12. Extracting features {F_{fi}} in RD_k by using the Rules R1-R1 based on opinion words EO={$OP_{1\text{-}i}$, $OP_{1\text{-}j}$ }; fi++;
13. **if** (Extracted new features not in F)
14. Extracting new features {F_{fj}} using Rules R31-R33 based on the new extracted features {F_{fi}}; fj++;
15. Extracting and updating new opinion words {$OP_{1\text{-}p}$} using Rules R21-R23 based on

extracted features F={F_{fi}, F_{fj} };
16. Extracting new features {F_{fp}} in RDk by using the Rules R1 based on new opinion words EO={$OP_{1\text{-}p}$}; fp++;
17. **End if**
18. Setting F={F_{fi}, F_{fj} , F_{fp} }; EO={$OP_{1\text{-}i}$, $OP_{1\text{-}j}$, $OP_{1\text{-}p}$ };
19. KernelFeature_OpinionSets=Build_kernel(F, EO, RDk);
20. Recording appearing frequency af of EO based on related F;
21. **if** The opinion words EO have the corresponding degree intensifier ODI
22. Building triple {ODI, EO, F}
23. **Else if**
24. Building triple {null, EO, F}
25. **End if**
26. Unique and update {ODI,EO,F};
27. Calculating the opinion polarity{OW} based on Definition 3.2.1- 3.2.3, Triple {ODI, EO, F}, and af;
28. **End for**
END

In order to test the proposed algorithm, the raw customer opinion data were collected by using publicly available information from the Amazon site. The experiments were conducted in three domains that including Canon camera, Casio watch, and Nike shoes. The test data included 3,458 customer reviews of 17 different type canon cameras, 354 customer reviews of Casio G-Shock watch, and 252 customer reviews of Nike woman shoes. Feature-by-feature comparison of the studied products is conducted based on the extracted features, opinions, and feature-opinion relations. Moreover, the strengths and the weaknesses of the studied products are given, which has a beneficial effect on the new product development and customer personalized recommendation.

5. The Extracted Results Comparison for the Proposed Two Approaches

In order to demonstrate the differences between the proposed two approaches, we analyzed 517 reviews about the product of Canon PowerShot SX280. The focused keywords are ‘zoom’, ‘video’, and ‘battery’. The TF-IDF approach can obtain the weight of each word in each document. The weights of studied keywords in seven documents are shown in Table 4. The results reveal the facts that the first document has the highest probability relevant with ‘zoom’ and the seventh document has the highest probability relevant with ‘video’. The keyword ‘battery’ appears in most of the documents, which means the majority of customers have discussed the attributes related to ‘battery’.

Table 4. Sample Output of the TF-IDF Approach

Document No.	1	2	3	4	5	6	7
Weight(‘zoom’)	0.1100	0	0.0204	0.0766	0	0.0464	0
Weight(‘video’)	0.0271	0	0.0151	0.0189	0	0.0229	0.0216
Weight(‘battery’)	0	0.0842	0.0197	0.0443	0.1010	0.0089	0.0112

The proposed opinion mining algorithm is adopted to extract information from the reviews related to 'zoom', 'video', and 'battery'. The important obtained information is analyzed in the following. More than half of the obtained feature-opinion in the battery dimension referred to a terrible battery quality, such as: 'bad battery life', 'battery died', 'battery drains', 'disappointed battery', and 'battery indicator issue' (Top 5 extracted negative frequent terms). Negative frequency terms of extracted feature opinion from the proposed algorithm are 'video problem', 'video issues', 'video shuts off', 'video not work', and 'disappointed video performance'. Moreover, 39.84% of extracted terms point to 'short battery life' and 21.48% are obtained as 'battery indicators issues'. Battery indicator issues are mainly about the indicators misleading the actual state of charge of the battery. The results also have 57 terms like 'defective firmware upgrade' in battery dimension. Therefore, we report the poor battery dimension, because of the battery life and the indicator problem; and the proposed solution from the company cannot completely solve the problem. As for the video dimension, the extracted results are more inconsistent and disorganized, such as: 'video camera died', 'minutes video battery shut(s) off', and 'zoom video mode battery shut down'. We can deduce that the video problem is probably caused by a battery problem.

Based on above analysis, the TF-IDF approach can be adopted to obtain the weights of studied keywords in reviews. The proposed opinion mining algorithm complements TF-IDF approach, which can extract more efficient information based on the content of reviews.

6. Conclusion

Information extraction is a tough problem that the existing approaches cannot obtain the desired extraction results. This paper globally views the existing approaches and then categorizes them into several groups based on the superiority and intelligence of the approaches and their capability to solve the complex information extraction (retrieval) problems. Two practical approaches are provided to demonstrate how to use the IE solutions based on different objectives. The first application aims to provide the front-end services for EASY-IMP project based on TF-IDF approach. The TF-IDF approach is adopted to analyze the customer's preference and determine the optimum interest of customers. TF-IDF approach is used to discover the most relevant keywords for the defined topics. Finally, roughly 100,000 pages have been analyzed and the customer's preference is determined based on the sets of selected keywords. In order to be more efficient for extracting the useful information from customer reviews, the opinion mining extraction algorithm is proposed. This algorithm can jointly identify features, opinion expressions, and feature-opinion, which capable to determine opinion boundaries and adopt syntactic parsing to learn and infer propagation rules between opinions and features. The proposed algorithm allows opinion extraction to be executed at the phrase level and can automatically detect the features that contain more than one word by building kernels through closest words. Experimental evaluations are conducted in 3,458 reviews and show that the proposed algorithm can complete the expected IE tasks. In the future, we will concentrate on testing the proposed algorithm. In order to obtain more accurate and efficient results, the proposed algorithm is considered as a supplement of TF-IDF approach when extracting information from various reviews.

References

1. Aggarwal, C.C., Zhai, C. (Eds.), 2012. Mining Text Data. Springer US, Boston, MA.
2. Aizawa, A., 2003. An information-theoretic perspective of tf–idf measures. Inf. Process. Manag. 39, 45–65.
3. Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.
4. Blunsom, P., 2004. Hidden markov models. Lect. Notes August 15, 18–19.
5. De Marneffe, M.-C., MacCartney, B., Manning, C.D., others, 2006. Generating typed dependency parses from phrase structure parses, in: Proceedings of LREC. pp. 449–454.
6. Dey, L., Verma, I., 2013. Text-Driven Multi-structured Data Analytics for Enterprise Intelligence. IEEE, pp. 213–220. doi:10.1109/WI-IAT.2013.186
7. Hiemstra, D., 2000. A probabilistic justification for using tf$\times$ idf term weighting in information retrieval. Int. J. Digit. Libr. 3, 131–139.
8. Hu, M., Liu, B., 2004. Mining and summarizing customer reviews, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 168–177.
9. Litvak, M., Last, M., 2008. Graph-based keyword extraction for single-document summarization, in: Proceedings of the Workshop on Multi-Source Multilingual Information Extraction and Summarization. Association for Computational Linguistics, pp. 17–24.
10. Liu, B., Zhang, L., 2012. A survey of opinion mining and sentiment analysis, in: Mining Text Data. Springer, pp. 415–463.
11. Li, Y., Krishnamurthy, R., Raghavan, S., Vaithyanathan, S., Jagadish, H.V., 2008. Regular expression learning for information extraction, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 21–30.
12. McCallum, A., 2005. Information extraction: Distilling structured data from unstructured text. Queue 3, 48–57.
13. McCallum, A., Freitag, D., Pereira, F.C., 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. In: ICML. pp. 591–598.
14. McDonald, R., Pereira, F., Ribarov, K., Hajič, J., 2005. Non-projective dependency parsing using spanning tree algorithms, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 523–530.
15. Nadeau, D., Sekine, S., 2007. A survey of named entity recognition and classification. Lingvisticae Investig. 30, 3–26.
16. Nivre, J., 2005. Dependency grammar and dependency parsing. MSI Rep. 5133, 1–32.
17. Paltoglou, G., Thelwall, M., 2010. A study of information retrieval weighting schemes for sentiment analysis, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 1386–1395.
18. Salton, G., Fox, E.A., Wu, H., 1983. Extended Boolean information retrieval. Commun. ACM 26, 1022–1036.
19. Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J., 2005. Developing a robust part-of-speech tagger for biomedical text, in: Advances in Informatics. Springer, pp. 382–392.