# Similarity-Based Classification for Big Non-Structured and Semi-Structured Recipe Data

Wei Chen[1,3] and Xiangyu Zhao[2,4(✉)]

[1] Agricultural Information Institute,
Chinese Academy of Agricultural Sciences, Beijing, China
[2] Beijing Research Center for Information Technology in Agriculture, Beijing, China
[3] Key Laboratory of Agri-information Service Technology,
Ministry of Agriculture, Beijing, China
chenwei@caas.cn
[4] National Engineering Research Center for Information Technology
in Agriculture, Beijing, China
zhaoxy@nercita.org.cn

**Abstract.** In current big data era, there has been an explosive growth of various data. Most of these large volume of data are non-structured or semi-structured (e.g., tweets, weibos or blogs), which are difficult to be managed and organized. Therefore, an effective and efficient classification algorithm for such data is essential and critical. In this article, we focus on a specific kind of non-structured/semi-structured data in our daily life: recipe data. Furthermore, we propose the document model and similarity-based classification algorithm for big non-structured and semi-structured recipe data. By adopting the proposed algorithm and system, we conduct the experimental study on a real-world dataset. The results of experiment study verify the effectiveness of the proposed approach and framework.

**Keywords:** Recipe data · Classification · User-generated contents · Semi-structured data · Non-structured data

## 1 Introduction

In current big data era, there has been an explosive growth of various data. Most of these large volume of data are non-structured or semi-structured (e.g., tweets, weibos or blogs). These data has the following distinct characteristics, which cannot be handled by the conventional database management systems.

- **Huge Volumed.** The user-generated data is large scale as users are easily to produce and share data with various kinds of devices.
- **Explosive Grown.** Every day, 2.5 quintillion bytes of data are produced and 90 % of the data in the world today were created within the past two years [20].


H. Gao et al. (Eds.): DASFAA 2016 Workshops, LNCS 9645, pp. 57–64, 2016.
DOI: 10.1007/978-3-319-32055-7_5

– **Non-/Semi-Structured.** The structure of the user-generated data are normally non-structured and semi-structured, since a piece of data (e.g., a tweet or a blog) may contain videos, images, texts and so on.

Therefore, it is very difficult to manage and organize such data by using the conventional approaches. To find the underline behaviors and patterns of the data, it is quite important to classify data into categories. To achieve this goal, an effective and efficient classification algorithm for the data is essential and critical. In this article, we focus on a specific kind of non-structured/semi-structured data in our daily life: recipe data. As the recipe data is a useful source to help people cook dishes, it is quite necessary to categorize these data to help them find favorite dishes of their preference. The main contribution of this paper are listed as follows.

– We present a document and similarity model for the big non-structured and semi-structured recipe data.
– We propose a similarity-based classification algorithm based on the document and similarity model.
– We conduct the experimental study on a real-world dataset, and verify the effectiveness of the proposed algorithm and framework.

The remaining parts of this article are organized as follows. Section 2 reviews the related work to our research. The proposed document and similarity modeling approach and the classification algorithm for big non-structured and semi-structured recipe data is introduced in Sect. 3. In Sect. 4, we report the experimental results by performing the algorithm and baseline in a real-world dataset. Section 5 summarizes the findings and some potential future directions of this research.

## 2 Related Work

In this section, we mainly review the research on the user-generated contents (e.g., twitters) in the Web 2.0 era. Golder and Huberman investigated the usage patterns and user behaviors of the social media contents and annotations [7]. Bischoff et al. did some statistical analysis on some social tagging data sets to gain valuable tags for search [2]. Manish et al. investigated various distinct characteristics of the user-generated tags [9]. Furthermore, the social tags were considered as an important semantic sources for modeling the items for recommendations [17,29], semantic retrieval [25] and personalized search [3,22]. More recently, researchers attempted to organize and categorize these data from the perspective of users. For example, the community-based modeling approaches [19,21,23,24] were adopted to achieve this goal. Another example is to identify the underlying patterns from the structure of social data [1,11]. Some researches also try to index the data through a cognitive approach, which identify sentiments [14,15], emotions [18], pre-knowledge [6,30], role in the group [5], preferred patterns [31] and personality [8]. There have been several classifying and organizing approaches for non-structured/semi-structured data. A classifier for the

semi-structured documents was proposed by using the a structured vector model [28]. Lesbegueries et al. presented models take into account characteristics of heterogeneous human expression modes: written language and captures of drawings, maps, pictures, etc., and semantic treatments have been built to automatically manage spatial and temporal information from non-structured data [13]. The techniques and relevant issues of keyword search were discussed in [4]. Mansmann et al. proposed an approach based on introducing a data enrichment layer responsible for detecting new structural elements in the data using data mining and other techniques [16]. EsdRank, which treats vocabularies, terms and entities from external data as objects connecting query and documents, was a new technique for improving ranking using external semi-structured data such as controlled vocabularies and knowledge bases [26].

## 3 Methodology

The overall workflow of classification algorithm is shown in Fig. 1. Firstly, the non-/semi-structured data will be pre-processed. The detail steps include Chinese segmentation, summarization and weighted value assignment. Secondly, the similarity among different documents will be calculated. Finally, we exploit the document similarity for classification, and obtain the categorized data.

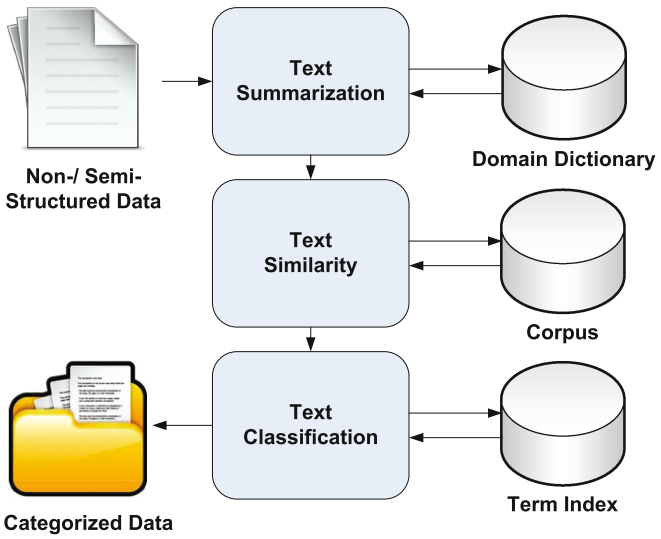


**Fig. 1.** The workflow of classification algorithm

### 3.1 Pre-process on Non/Semi-Structured Data

As the dataset is Chinese, we firstly need to segment Chinese texts into terms. To address this issue, we employ the Hidden Markov Model (HMM) to perform

the Chinese segmentation. The detail of our segmentation approach is introduced in [27]. Furthermore, we build a domain dictionary to extract the summary of the texts, filter trivial terms and assign the weight to these terms. The construction of domain dictionary follows the approach introduced in [10]. After the pre-processed, we then attempt to construct the relationship among documents.

### 3.2 Document and Similarity Modeling

By pre-processing the texts, we obtain a set of documents to be categorized. Formally, we defined a document as

$$d_i =< t_1 : w_i(t_1); t_2 : w_i(t_2); ...t_n : w_i(t_n) > \tag{1}$$

where $t_x$ is a term of the document, and $w_i(t_x)$ is the weight to be assigned to the term. $w_i(t_x)$ is calculated based on the hybrid paradigm [22] as follows.

$$w_i(t_x) = f_i(t_x) \times \frac{log^2 N}{m_x^2} \times \frac{f_i(t_x) \cdot (k+1)}{f_i(t_x) + k \cdot (1 - b + b \cdot \frac{l_i}{l_a})} \tag{2}$$

where $f_i(t_x)$ is the term frequency of $t_x$ in the document, $N$ is the total number of documents, $m_x$ is the number of document containing $t_x$, $l_i$ is the length of the document, $l_a$ is the average length of all documents, and $k$ and $b$ are two parameters, which are set as 2 and 0.75.

By modeling the document, we obtain a vector for each document. It is quite straightforward to adopt the cosine measurement to compute similarity between documents.

$$Sim(d_i, d_j) = \frac{\boldsymbol{d_i} \cdot \boldsymbol{d_j}}{||\boldsymbol{d_i}||||\boldsymbol{d_j}||} \tag{3}$$

Therefore, we can construct the similarity between each pair of documents in the corpus.

### 3.3 Classification Based on Similarity

The algorithm of classification is based on the similarity we have obtained in the last step. As shown in Algorithm 1, the core idea is adapted from the prototype-based clustering [12]. Firstly, we randomly assign documents to each category as a prototype. Secondly, we assign the document has the maximal similarity with the prototype to the category. Finally, the class labels are outputs when all documents in the corpus has been assigned.

## 4 Experiment

The experiment is based on a dataset collected from a folksonomy-based multi-media recipe dataset in [3]. The dataset contains 500 recipes (documents) in five main kinds of dishes in China (i.e., Cantonese dishes, Sichuan dishes and so on)

**Data**: A set of documents $D$ ($d_i \in D$); The set of categories $C$ ($c_x \in C$)
**Result**: Class label $c_x$ for $d_i$
**for** *each* $c_x \in D$ **do**
  Randomly assign $|C|$ documents to each $c_x$;
  Set these documents $d_x$ as the prototype of $c_x$;
  $D \leftarrow D - \{d_x\}$
**end**
**for** *each document* $d_i \in D$ **do**
  Compute $sim(d_i, d_x)$ by Equation (3);
  Find the$d_i$,$d_x$ with maximal $sim(d_i, d_j)$;
  $c_x \leftarrow d_i$;
  $d_x \leftarrow Mean(c_x$;
  $D \leftarrow D - \{d_i\}$;
**end**
Output class label $c_x$ for $d_i$;

**Algorithm 1.** The Similarity-based Classification

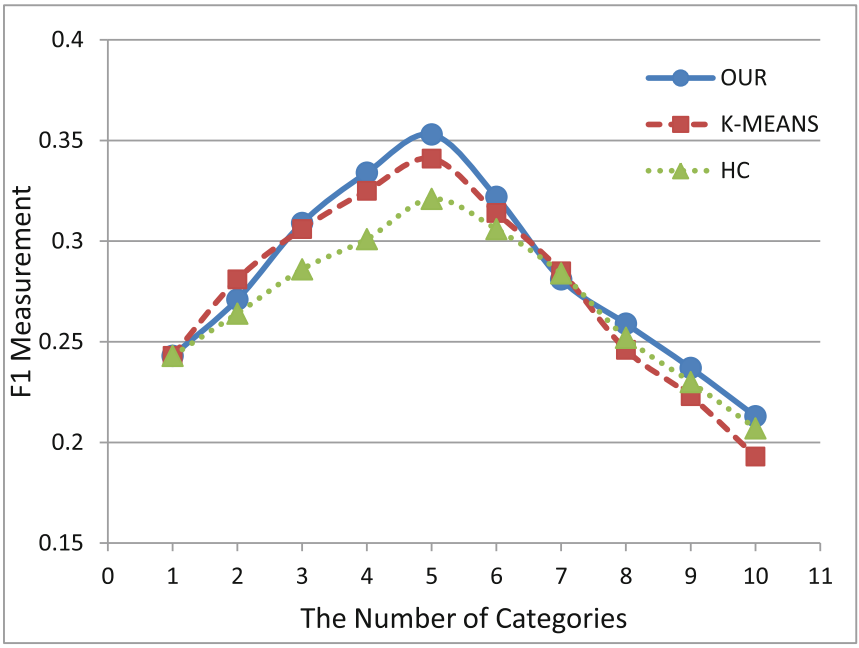


**Fig. 2.** The performance of F1 measurement

associated with the general descriptions on the characteristics, ingredients, main taste and so on. Furthermore, 203 users have annotated 7,889 tags on these recipes. Averagely, each user has annotated 16.7 recipes. For the purpose of evaluating the classification algorithm on non/semi-structured documents, we combine the user-generated tags for each recipe with the corresponding recipe descriptions to generate a non/semi-structured document.

To evaluate the effectiveness of the proposed framework and algorithm, we use the F1 measurement, which is the harmonic mean of precision and recall. As the proposed method is an unsupervised method for classification, we adopt a voting strategy to determine the predicted classification label in a category. That is, the final label of a category is determined by the label which have the most members in the category. The accuracy and recall are therefore evaluated by matching the final labels and the ground-truth labels. As shown in Fig. 2, the performance of the proposed methods (OUR), K-Means (K-MEAN) and Hierarchial method (HC) are illustrated. It is also quite clear that the F1 measurement achieve by our method has the best performance ($F1 = 0.535$, $c = 5$). The effectiveness of our proposed method is verified. It is worth to point out that all methods have the best performance in the number of categories ($c = 5$). A reasonable explanation for this observation is that the dataset contains five kinds of dishes as we mentioned. The classification is overfitting the dataset when the number of categories is greater than 5, while it is underfitting when the number of categories is less than 5.

## 5 Conclusion

In this paper, we present a document and similarity model for the big non-structured and semi-structured user-generated data. Based on the model, we propose a similarity-based classification algorithm for processing the user-generated data. Furthermore, We conduct the experimental study on a real-world dataset, and verify the effectiveness of the proposed algorithm and framework. In future research, we plan to build user profiles to further improve the effectiveness of the system.

**Acknowledgement.** This work is supported by Fundamental Research Funds of Agricultural Information Institute, Chinese Academy of Agricultural Sciences (No. 2014-J-011), and Project of Ministry of Agriculture of China "Agricultural information monitoring and early-warning".

## References

1. Armstrong, T.G., Ponnekanti, V., Borthakur, D., Callaghan, M.: Linkbench: A database benchmark based on the facebook social graph. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 1185–1196. ACM (2013)
2. Bischoff, K., Firan, C.S., Nejdl, W., Paiu, R.: Can all tagsbe used for search? In: Proceedings of CIKM 08, Napa Valley, California, USA, October 26-30, pp. 193–202. ACM, New York, NY, USA (2008)
3. Cai, Y., Li, Q., Xie, H., Yu, L.: Personalized resource search by tag-based user profile and resource profile. In: Chen, L., Triantafillou, P., Suel, T. (eds.) WISE 2010. LNCS, vol. 6488, pp. 510–523. Springer, Heidelberg (2010)

4. Chen, Y., Wang, W., Liu, Z., Lin, X.: Keyword search on structured and semi-structured data. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 1005–1010. ACM (2009)
5. Feng, X., Peng, Y., Xie, H., Yan, Z.: Role-based learning path discovery for collaborative business environment. In: International Conference on Control, Automation and Systems Engineering (CASE), pp. 1–4. IEEE (2011)
6. Feng, X., Xie, H., Peng, Y., Chen, W., Sun, H.: Groupized learning path discovery based on member profile. In: Luo, X., Cao, Y., Yang, B., Liu, J., Ye, F. (eds.) ICWL 2010. LNCS, vol. 6537, pp. 301–310. Springer, Heidelberg (2011)
7. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. J. Inf. Sci. **32**, 198–208 (2006)
8. Gou, L., Zhou, M.X., Yang, H., Knowme, S.: Understanding automatically discovered personality traits from social media and user sharing preferences. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 955–964. ACM (2014)
9. Gupta, M., Li, R., Yin, Z., Han, J.: Survey on social tagging techniques. SIGKDD Explor. Newsl. **12**, 58–72 (2010)
10. Islam, A., Inkpen, D.: Semantic text similarity using corpus-based word similarity and string similarity. ACM Trans. Knowl. Disc. Data (TKDD) **2**(2), 10 (2008)
11. Jin, T., Xie, H., Lei, J., Li, Q., Li, X., Mao, X., Rao, Y.: Finding dominating set from verbal contextual graph for personalized search in folksonomy. In: IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 1, pp. 367–372. IEEE (2013)
12. Kuncheva, L., Bezdek, J.C., et al.: Nearest prototype classification: Clustering, genetic algorithms, or random search? IEEE Trans. Syst. Man Cybern., Part C: Appl. Rev. **28**(1), 160–164 (1998)
13. Lesbegueries, J., Gaio, M., Loustau, P.: Geographical information access for non-structured data. In: Proceedings of the ACM Symposium on Applied Computing, pp. 83–89. ACM (2006)
14. Li, X., Xie, H., Chen, L., Wang, J., Deng, X.: News impact on stock price return via sentiment analysis. Knowl. Based Syst. **69**, 14–23 (2014)
15. Li, X., Xie, H., Song, Y., Li, Q., Shanfeng Zhu, F., Wang, L.: Does summarization help stock prediction? News impact analysis via summarization. IEEE Intell. Syst. **30**, 26–34 (2015)
16. Mansmann, S., Rehman, N.U., Weiler, A., Scholl, M.H.: Discovering olap dimensions in semi-structured data. Inf. Syst. **44**, 120–133 (2014)
17. Mao, X., Li, Q., Xie, H., Rao, Y.: Popularity tendency analysis of ranking-oriented collaborative filtering from the perspective of loss function. In: Bhowmick, S.S., Dyreson, C.E., Jensen, C.S., Lee, M.L., Muliantara, A., Thalheim, B. (eds.) DASFAA 2014, Part I. LNCS, vol. 8421, pp. 451–465. Springer, Heidelberg (2014)
18. Rao, Y., Lei, J., Wenyin, L., Li, Q., Chen, M.: Building emotional dictionary for sentiment analysis of online news. World Wide Web **17**(4), 723–742 (2014)
19. Tang, J., Chang, Y., Liu, H.: Mining social media with social theories: A survey. ACM SIGKDD Explorations Newsletter **15**(2), 20–29 (2014)
20. Xindong, W., Zhu, X., Gong-Qing, W., Ding, W.: Data mining with big data. IEEE Trans. Knowl. Data Eng. **26**(1), 97–107 (2014)
21. Xie, H.-R., Li, Q., Cai, Y.: Community-aware resource profiling for personalized search in folksonomy. J. Comput. Sci. Technol. **27**(3), 599–610 (2012)
22. Xie, H., Li, Q., Mao, X.: Context-aware personalized search based on user and resource profiles in folksonomies. In: Sheng, Q.Z., Wang, G., Jensen, C.S., Xu, G. (eds.) APWeb 2012. LNCS, vol. 7235, pp. 97–108. Springer, Heidelberg (2012)

23. Xie, H., Li, Q., Mao, X., Li, X., Cai, Y., Rao, Y.: Community-aware user profile enrichment in folksonomy. Neural Netw. **58**, 111–121 (2014)
24. Xie, H., Li, Q., Mao, X., Li, X., Cai, Y., Zheng, Q.: Mining latent user community for tag-based and content-based search in social media. Comput. J. **57**(9), 1415–1430 (2014)
25. Xie, H., Yu, L., Li, Q.: A hybrid semantic item model for recipe search by example. In: IEEE International Symposium on Multimedia (ISM), pp. 254–259. IEEE (2010)
26. Xiong, C., Callan, J.: Esdrank: Connecting query and documents through external semi-structured data. In: International Conference on Information and Knowledge Management, pp. 951–960. ACM (2015)
27. Yang, W., Ren, L.-Y., Tang, R.: A dictionary mechanism for chinese word segmentation based on the finite automata. In: International Conference on Asian Language Processing (IALP), pp. 39–42. IEEE (2010)
28. Yi, J., Sundaresan, N.: A classifier for semi-structured documents. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 340–344. ACM (2000)
29. Yu, L., Li, Q., Xie, H., Cai, Y.: Exploring folksonomy and cooking procedures to boost cooking recipe recommendation. In: Du, X., Fan, W., Wang, J., Peng, Z., Sharaf, M.A. (eds.) APWeb 2011. LNCS, vol. 6612, pp. 119–130. Springer, Heidelberg (2011)
30. Zou, D., Xie, H., Li, Q., Wang, F.L., Chen, W.: The load-based learner profile for incidental word learning task generation. In: Popescu, E., Lau, R.W.H., Pata, K., Leung, H., Laanpere, M. (eds.) ICWL 2014. LNCS, vol. 8613, pp. 190–200. Springer, Heidelberg (2014)
31. Zou, D., Xie, H., Wang, F.L., Wong, T.-L., Wu, Q.: Investigating the effectiveness of the uses of electronic and paper-based dictionaries in promoting incidental word learning. In: Cheung, S.K.S., Kwok, L.-F., Yang, H., Fong, J., Kwan, R. (eds.) ICHL 2015. LNCS, vol. 9167, pp. 59–69. Springer, Heidelberg (2015)