Improving item to item text similarity for content-based recommender systems in collaborative question answering sites using a clustering ensemble approach

Short Paper

ABSTRACT

Collaborative Question Answering (CQA) sites allow community members to ask and answer questions, based on their expertise. Taking advantage of the rich resources (i.e., questions and answers) already available on a CQA site, we present an clustering ensemble approach for text comparison with application in question recommendation, in order to reduce the time and effort required to obtain an appropriate response for community questions. We propose a novel proximity measure which improves the content based recommendation scheme by providing an adimensional and unbiased value that represents effectively the underlying structure for text items. This allows the CQA site to enhance user experience by recommending related text items in terms of meaning and content simultaneously. Given the challenges involved in deploying such a strategy in real time, we also offer some insights regarding overcoming restrictions on volume and velocity that are part of a Big Data driven context. We have conducted initial assessments using the Quora dataset, obtaining promising results in terms of recommendation improvement over baselines.

KEYWORDS

Content-based recommendation, Text similarity, Ensemble clustering, Collaborative question answering

1 INTRODUCTION

Collaborative Question Answering (CQA) sites gather community members appealing to their needs to ask and answer questions, based on their expertise [10]. Popular sites with this purpose, such as *Yahoo! Answers*, *StackOverflow* or *Quora*, are accessed every day by millions of users[1]. One challenge regarding these sites is the time required for a user to obtain a proper answer for a given question. A well-known approach regarding this issue is to find a similar question within the site database that may have been asked beforehand, based on the premise that its respective answer might also be appropriate for the original user question [14]. Content-based recommender systems are able to tackle this issue by applying techniques that select the most related item for a question provided as input [2]. One key feature for the effectiveness of a recommender system is the ability of representing the underlying similarity among items in an appropriate manner [1]. In a CQA site, items -questions and answers- can be represented as text objects. These type of data convey particular issues when compared quantitatively, in terms of sentence content and meaning. Several measures were proposed to find values that appropriately express relationships among pieces of text, such as those based on weighted frequencies of words [5, 11, 12]. These measures are based on several aspects of the words, n-grams, or sentences. In this paper we present a novel strategy that leverages text analysis to recommend questions, in order to reduce waiting time and provide meaningful answers. We propose a novel measure based on an clustering ensemble approach [6]. Our measure consider the co-association between elements across several series of clustering runs. Each of these series is, in turn, based on a state of the art text similarity measure. The result is an adimensional and unbiased value that can improve the representation for the underlying structure of text relationships. In addition to this, some insights were found on the challenges involved in the application of our strategy in real time, regarding overcoming restrictions on volume and velocity that are part of a Big Data driven context. We have performed initial assessments on a large dataset from Quora, obtaining promising results over baselines. Therefore, our clustering ensemble approach can be incorporated in a content based recommender scheme in order to provide effective and reliable results for finding related questions in CQA sites. This work is structured as follows. In the Clustering ensemble approach section, we explain our method to obtain an adimensional and unbiased text proximity measure. In the Experiments and results section, we present the initial assessments on the Quora dataset, and offer insights on the implementation of the method in a real time environment. In the Discussion section, we provide some conclusions on the proposal and results obtained with our method.

2 CLUSTERING ENSEMBLE APPROACH

In this section, we present our approach for improving content based recommendation systems for CQA sites. This method relies upon obtaining an integrated means of representing the underlying structure of text item relationships. It is based on clustering ensembles through evidence accumulation, as stated in [6]. The details of the method are described below.

2.1 Clustering ensembles

A clustering algorithm aims to find an appropriate grouping structure for a given set of objects, in the form of a data partition. It is known that the use of different clustering algorithms, or even the same algorithm with different parameters, produce different results for the same input data [15]. In the case of text objects such as questions and answers, where the original words and sentences

[1] A question asked in Quora, and answered by its CEO and founder Adam D'Angelo reveals that the site receives over 200 million monthly unique visitors. http://quora.com/How-many-people-use-Quora-7. Last Accesed May 2018.

RecSys18, 12th ACM Conference on Recommender Systems, Vancouver, Canada

cannot be located in a dimensional space, there is a known variation in results according to the structural aspects taken into account by the input proximity measure [7]. Considering this situation, the evidence accumulation approach combines a set of different resulting partitions into a clustering ensemble that summarizes the relations between data through all the input clustering algorithms and/or parameters. Intuitively, if a pair of text objects are closely related, in a way that they are very similar to each other in terms of several structural aspects such as content and meaning, most of the clustering algorithms applied to them will co-cluster both within the same group. Thus, through different partitions created from the same original data, similar text objects are likely to belong to the same cluster. Based on this premise, the clustering ensemble approach aims to perform different input algorithms, in order to obtain a diverse set of resulting partitions. Then, a co-clustering assessment is performed for each pair of objects. The proportion of times that a pair of text objects appear in the same group over the complete set of partitions is calculated. This result is a value in the range $[0, 1]$ that constitutes a new distance measure for the original data. It considers the likelihood that two text objects are clustered together when different input clustering algorithms are applied. If the proportion of times is close to 1, it is clear that both text objects share several aspects that make them similar in terms of the true underlying structure of the data, and vice-versa. Because of how this calculation is performed, this result is an adimensional and non-biased value that comprises the variability of the input clustering algorithms. This new measure serves as an integrated representation of the underlying relationships within the original data. Therefore, the recommendation task is effectively improved by obtaining item to item text distances in a highly realistic manner, in an effort to better describe the real structure of the data.

2.2 A workflow for the clustering ensemble approach

The workflow presented for our method consists of two steps, as shown in Figure 1. The first step aims to obtain a diverse set of text object partitions. This is generated by applying clustering algorithms on the original data taking into account a reasonable degree of random initialization. This is obtained by variating a) the input proximity measure, considering different aspects of the text objects, and b) the final number of clusters k, selected from a restricted range, as proposed in [6]. This step generates a diverse partition set, that summarizes the different clustering structure through all the individual partitions. In the second step, a co-clustering value is calculated to evaluate the agreement between all pairs of objects. These agreement values are then disposed in a co-association matrix, as a way of representing the pairwise distances. This matrix constitutes an improved item to item text distance structure to support the content based recommendation scheme.

2.3 Generating a diverse set of partitions

In the first step, the diverse set is created based on a clustering algorithm. Since some text distance definitions might not be represented in a dimensional space [9], an algorithm that is able to use a distance matrix directly as input is preferred [8]. Therefore, given an input data set of size n, a number D of text distance matrices for the n text objects are used as input. For each distance matrix, a number N of values k is selected randomly within a restricted range. The k value is the input clustering parameter for the resulting number of clusters. This combination of D distance matrices and N number of clusters for each distance matrix yields a total of $D \times N$ clustering runs. This configuration aims to ensure a degree of variability for the output partitions. Finally, a diverse partition set $\mathbb{P}$ is built to summarize the structure of the clustering algorithms through all the partitions, $\mathbb{P} = \{P_{11}, P_{12}, \ldots, P_{DN}\}$, where P_{ij} is a partition for the distance matrix $i, 1 \leq i \leq D$ combined with the input k parameter for the j-th clustering run, $1 \leq j \leq N$. In turn, each partition P_{ij} contains pairs of elements indicating a cluster identifier $C_p, 1 \leq p \leq k$ and the text object assigned to that cluster $T_q, 1 \leq q \leq n$. Therefore $P_{ij} = \{(C_1, T_1), (C_1, T_2), \ldots (C_k, T_n)\}$

The diverse partition set $\mathbb{P}$ provides compact information on the assigned cluster for each object over all the clustering runs. This information will be used as an input to the next step, as detailed below.

2.4 Building a co-association proximity matrix

Based on the diverse partition set, an assessment on the co-clustering for each pair of objects is performed. With the information of the clustering assignment, an evaluation of the proportion of times that a pair of objects i, j calculated as a co-association index $c_{i,j} = \frac{n_{(i,j)}}{DN}$, where $n_{(i,j)}$ is the number of times that the pair of text objects (i, j) is assigned to the same cluster over the $D \times N$ partitions from the diverse partition set. This is a similarity measure that can be easily converted to a distance measure by calculating $1 - c_{i,j}$. This value is then assigned to the corresponding position i, j within a co-association matrix $\mathbf{M}$, so that $\mathbf{M}_{ij} = 1 - c_{i,j}$. The co-association index is an adimensional, non biased measure that comprises the variability of the applied clustering algorithms, and serves as a proximity measure that represents the underlying structure of the text objects in an integrated way. Therefore, this co-association distance matrix enhances the item to item distance structure needed for a content based recommendation scheme by incorporating several aspects of the text distances rather than using a simple measure based on individual aspects such as the weighted frequencies of words.

3 EXPERIMENTS AND RESULTS

In this section, we describe the experiments performed with our approach on the Quora dataset. We explain the base text measures from the state of the art and the input parameters selected as input for the clustering algorithm. We also offer some insights on the application of our method in a real time, Big Data driven context. Finally, we present initial results obtained on the quality of the item to item distance measure for the content based recommender scheme.

3.1 Input data set

We used a dataset released from the Quora website [2]. This dataset consists of 404290 pairs of potential question duplicate pairs. An extract from the dataset is shown in Table 1.

Each pair contains a pair of questions written in common English language, with an indicator value that has a value of 1 for duplicated questions, and 0 for questions that are not duplicated. There are

[2] "First Quora Dataset Release: Question Pairs". Available in https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs. Last accesed May 2018.

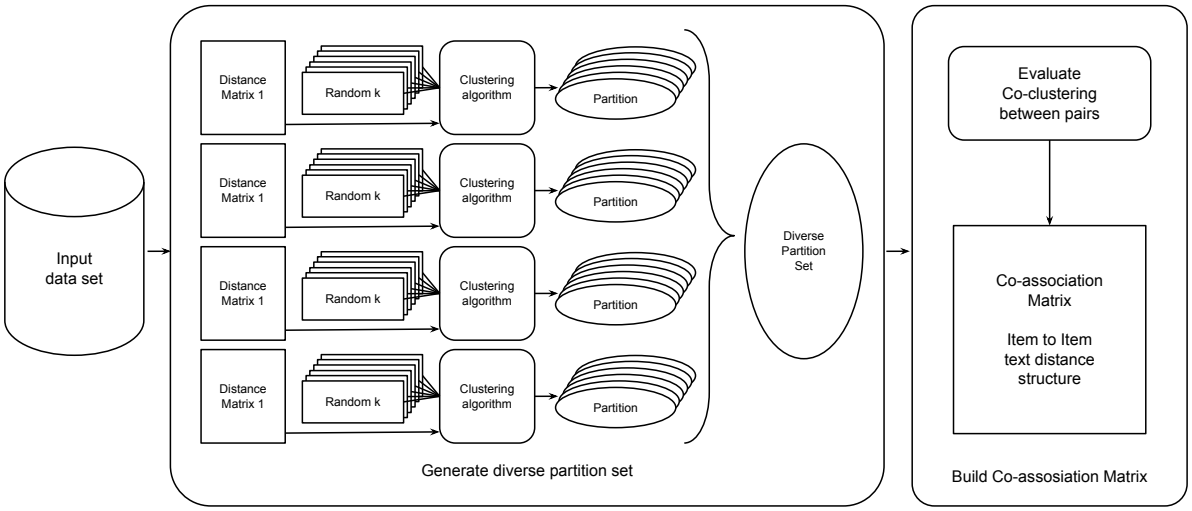

Figure 1: Workflow for the clustering ensemble approach.

Table 1: Extract from the Quora dataset

questionone	questiontwo	is_duplicate
a jellyfish can eat another jellyfish	how is a jellyfish able to eat	0
a long distance relationship does this work	do long distance relationships last	1
what is a good alternative for google	what are the best alternatives to google	1
what is a god	what is beyond god	0

some major challenges when dealing with a dataset of this size, as we will describe below. There is an imbalance in the duplicity of the question, with a total of 149.263 pairs labeled as duplicated (36,9%) and the remaining 255.027 pairs labeled as not duplicated (63,1%). Therefore, to avoid potential biases in the results, we decided to work with a number of samples with approximately balanced labeling. We generated 100 samples of size 100 (100 question pairs for each sample), with a proportion of duplicated questions between 40% and 60%. The value for the proportion was taken from a computationally generated uniform distribution in that percentage range.

3.2 Base text distance measures

The base text distance measures used in the experiment were selected following a preliminary text distance comparative analysis studied in [3]. These measures are briefly described below.

The first base measure used is *Term Frequency (TF)*. This is a well-known technique commonly used in the field of information retrieval [13]. In this technique, a cosine distance is calculated between two vectors that represent the number of times that the terms appear in two given pieces of text. The second base measure used is *TF/IDF*. This is a well-known distance commonly used in content based recommender schemes [4]. It consists on applying the TF technique with an additional weighting on the term frequencies, taking into account the inverse proportion of terms over the entire text corpus. The third measure used is *Word2Vec*. This is a series of models used to represent words as vectors based on their context [11]. Therefore, words with similar meaning have vectors that are close to each other. In particular, we use the *continuous bag-of-words* model, which predicts a center word given past and future words defined by a text window. The vector representation of the window words are averaged and a softmax function is performed to obtain the center word. The fourth measure used is *FastText*. This is a continuous word representation technique which takes into account the morphology of words [9]. FastText consists in a neural network trained to predict surrounding words in a sentence. The word is represented as a vector calculated as the sum of the vector representations of its character n-grams. The resulting vectors are then compared using the cosine distance mentioned above.

3.3 Diverse set of partitions

The diverse set was generated as follows. Since the Quora dataset was structured into question pairs, in the form *(question1,question2)*, for each sample of size 100 the individual questions were extracted and re-ordered to conform a list of $n = 200$ unique text objects. Then, the four distances mentioned above were calculated pairwise for all questions. For each distance, $N = 100$ values of k were selected randomly in the range $[5, 10]$ as suggested in [6]. Using these parameters as input, the commonly used PAM clustering algorithm was applied [8], with random initialization of the medoid objects. This process was performed over the 100 samples to obtain the corresponding diverse sets of partitions. Each set consisted of $D \times N = 400$ partitions. Each of these partitions contained the clustering assignment information for the original 200 text objects.

3.4 Co-association matrix

Based on the obtained diverse partition sets, a co-association matrix was generated by calculating the co-association index over all the partitions within each partition set. Therefore, a total number of 100 distance matrices was generated with the clustering ensemble approach. These distances were compared with the real data, in terms of the indicator flag, using a classic machine learning approach to define the training, validation and test sets. A n-fold cross validation framework was used, in a way that n training / validation sets were

Table 2: Confusion matrix output for the base distances and the clustering ensemble approach

			Predicted	
			0	1
TF	Real	0	0.38	0.16
		1	0.10	0.36
TF/IDF	Real	0	0.38	0.16
		1	0.09	0.37
Word2Vec	Real	0	0.35	0.19
		1	0.11	0.35
FastText	Real	0	0.42	0.12
		1	0.26	0.20
Clustering ensemble	Real	0	0.39	0.15
		1	0.18	0.28

defined. A portion of the data was withheld for the final testing. Once these sets were defined, the clustering ensemble approach was assessed for classification performance. For each clustering run, a threshold value was selected from the training set taking into account the one that yielded the smallest error for the validation set. These error values were then averaged over the total number of runs for each sample.

3.5 Results

In this section, we present the results obtained from the performed experiments. Table 2 shows the confusion matrices for all the base distances, along with the clustering ensemble approach. Each row shows the validation results for each distance measure. Percentual values for the confusion matrices are presented on each column. In each confusion matrix the output classes are presented with the value of 0 for non duplicated questions and 1 for duplicated questions. For each measure, the real class is displayed on rows, whereas the predicted class is displayed on columns.

In the table, it can be seen that the clustering ensemble approach yields an excellent overall performance when compared with the base distance methods. It has a high quality classification performance for predicting different items (with a value of 0.39) when compared with the rest of the base distances, except FastText, which has a value of 0.42. However, the clustering ensemble approach has a better performance regarding prediction of non different questions than FastText, with a value for the clustering ensemble approach of 0.28 against a lower value of 0.20 for FastText. Moreover, although the error value of 0.15 for predicting different questions for our method is slightly higher than the FastText value of 0.12, the complementary error rate for predicting non different questions in the clustering ensemble approach with the value of 0.18 is considerably lower than the respective value of 0.26 for FastText. When compared with the rest of the base distances other than FastText, the clustering ensemble approach has also the lowest error rate for predicting different questions. The overall results shows that the clustering ensemble approach predicts different text objects with high quality when compared to the base distances TF, TF/IDF and Word2Vec. As well as this, our approach predicts non different questions with high performance when compared to FastText. Therefore, our preliminary assessments with the clustering ensemble approach provide results that are promising for further development of our method.

3.6 Challenges in a Big Data context

It is necessary to highlight the challenges that were found in the performed experiments. The calculation of the input distances and the assessments of the results were developed in Python, whereas the main code for the ensemble approach was developed in R. Since our method requires working with a large input dataset, we encountered a series of constraints regarding processing time. In particular, the calculation time for non-trivial distances such as FastText and Word2Vec over the 100 samples was within the range of 3 to 6 hours in a 6-core standard processor. Once all distances were calculated, the clustering ensemble approach processing time is about 1 to 2 hours. However, considering the complete Quora dataset, the number of distances to be calculated is $\frac{n(n+1)}{2} = 326901212490$ values, which is a large value to be processed with a desktop computer. This leads the developer to consider a Big Data architecture, possibly optimizing the code by applying parallelization techniques over a large number of cores. Regarding storage space, it should be noted that a matrix those distances that require double precision for internal representation has a size of approximately 750KB when stored in a file. Hence, should the technique be applied to the complete dataset, it would require about 12TB for each distance matrix. Therefore, an optimized storage scheme should be considered for the effective implementation of the method.

4 DISCUSSION AND FUTURE WORK

We have proposed and tested a new method for improving the item to item distance structure required in a content based recommender scheme. This method aims to improve the results required within CQA sites regarding text objects. Our method proved to perform with high quality when compared to state of the art distances used for text objects comparison. In particular, out clustering ensemble approach yielded excellent results when compared to the commonly used TF/IDF measure, regarding the classification performance on pairs of questions that are different from each other. Future work is to be made in terms of working with larger sample sizes, and ultimately with the complete Quora dataset, or a similar real text dataset from a CQA website. In this direction, an appropriate Big Data architecture should be designed in order to tackle the challenges associated when dealing with a large complex text dataset. Our results constitute a solid starting point for building a reliable item to item distance structure in order to enhance the content based recommendation scheme applied to CQA sites. Therefore, this work opens a path for better studying the underlying structure on text data that define the real relationships between text items. With an improved method to assess this data structure, more reliable recommendations are able to be obtained upon a content based scheme, on the basis of better recognizing the complex true relations that build up the question and answer text datasets within the CQA sites.

REFERENCES

[1] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering* 17, 6 (2005), 734–749.

[2] Charu C Aggarwal. 2016. Content-based recommender systems. In *Recommender Systems*. Springer, 139–166.

[3] Franco Ferrari Guadalupe Guereta Mercedes Valoni Santiago Diez Sole Pera Ion Madrazo Azpiazu Guillermo Leale Alejandro Gonzalez, Carlos Flury. 2018. Comparative Analysis on Text Distance Measures Applied to Community Question Answering Data. *5to Congreso Nacional de Ingeniería Informática. Córdoba, Argentina.* (2018).
[4] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. *Modern information retrieval*. Vol. 463. ACM press New York.
[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* (2016).
[6] Ana L.N. Fred and Anil K. Jain. 2005. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 6 (jun 2005), 835–850. https://doi.org/10.1109/TPAMI.2005.113
[7] Andreas Hotho, Andreas Nürnberger, and Gerhard Paaß. 2005. A brief survey of text mining.. In *Ldv Forum*, Vol. 20. Citeseer, 19–62.
[8] Leonard Kaufman and Peter J Rousseeuw. 2009. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons.
[9] Yuhua Li, David McLean, Zuhair A Bandar, James D O'shea, and Keeley Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE transactions on knowledge and data engineering* 18, 8 (2006), 1138–1150.
[10] Yuanjie Liu, Shasha Li, Yunbo Cao, Chin-Yew Lin, Dingyi Han, and Yong Yu. 2008. Understanding and summarizing answers in community-based question answering services. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 497–504.
[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
[12] Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. In *The adaptive web*. Springer, 325–341.
[13] Amit Singhal et al. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.* 24, 4 (2001), 35–43.
[14] Kai Wang, Zhaoyan Ming, and Tat-Seng Chua. 2009. A syntactic tree matching approach to finding similar questions in community-based qa services. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 187–194.
[15] Rui Xu and Don Wunsch. 2008. *Clustering*. Vol. 10. John Wiley & Sons.