The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence
Knowledge Extraction from Text: Technical Report WS-16-10

Automatic Scoring for Innovativeness of Textual Ideas

Tirthankar Dasgupta and Lipika Dey
TCS Innovation Lab
{gupta.tirthankar, lipika.dey}@tcs.com

Abstract

Automatic evaluation of text for its innovative quality has been necessitated by the growing trend to organize open innovation contests by different organizations. Such online/offline contests are known to fuel major business benefits to many industries. However, open contests result in a huge number of documents of which only a few may contain potentially interesting and relevant ideas. Usually these entries are manually reviewed and scored by multiple experts. But manual evaluation process not only require a lot of time and effort but are also prone to erroneous judgments due to inter-annotator disagreements. To counter this issue, in this paper, we have proposed a new approach towards detecting novelty or innovativeness of textual ideas from a given collection of ideas. The proposed approach uses information theoretic measures and term relevance to domain to compute document level innovativeness score. We have evaluated the performance of the proposed approach with a real world collection of innovative ideas which were manually scored by experts. We have compared the performance of our proposed model with some of the commonly used baseline approaches that rely on distributional semantics and geometric distances. The result shows that the proposed method outperform the existing baseline models.

Introduction

Open innovation contests have become extremely popular in harnessing crowd-sourced ideas and are exploited by various organizations for their business benefits. The term open innovation refers to crowd-sourcing based online/offline contests that bring together customers, employees and partners to generate unrealized ideas that can eventually help the growth of the particular organization (Chesbrough, 2006). However, conducting such crowd-sourced contests results in a huge number of potential ideas that may or may not be relevant to the concerned organization. Therefore, these generated ideas are required to be carefully reviewed multiple times by experts manually. Consequently, such manual evaluation processes becomes expensive not only in terms of time and effort but also may be prone to erroneous judgments. Therefore, developing a means through which such ideas can be automatically scored to eliminate poor ideas can save a lot of valuable time and effort of the expert reviewers. However, organizations do not want to lose any single idea even if it is half-baked or not well-formed. The selection criteria needs to be accordingly designed.

This problem emerges in a plethora of other applications ranging from email thread filtering to RSS news feed recommendation and is commonly termed as Novelty/Innovativeness[1] Detection. Therefore, Novelty Detection is the task of assigning novelty score to a given document based on a set of already accumulated background information.

A number of attempts have been made to compute novelty of text documents. These, attempts have primarily focused on measuring the similarity distance between documents. (Allan, Wade, & Bolivar, 2003) (Allan, Lavrenko, Malin, & Swan, 2000) (Markou & Singh, 2003) (Lin & Brusilovsky, 2011)(Lin & Brusilovsky, 2011). However, none of the existing approaches have reported a satisfactory performance towards this end. Moreover, the existing methods involving the geometric distance and distributional semantics are extremely resource sensitive and performs poorly in resource constraint circumstances (Margarita, Rousseau, Ntoulas, & Vazirgiannis, 2014). The above facts clearly demonstrate that the novelty detection task is quite difficult to achieve under the current scenario and very much a subject of research.

The primary objective of this paper is to automatically score the novelty of an idea with respect to a collection of ideas. The solution is designed to keep the following issues in focus:

(i) An idea is scored higher for innovativeness if it contains a unique concept or technique that is not discussed in any other document. (ii) When an idea is not exactly unique but rare in the collection, then the content that presents the idea in a more comprehensive manner should be scored higher over others that contain similar idea.

[1] We have considered the terms novelty and innovativeness to be same and will use them interchangeably.

Accordingly, we have capitalized on the concept of information theory to determine the information content of a given piece of text. Our hypothesis is that a document having high information content is potentially a novel document. We have compared the performance of our proposed model with some of the commonly used baseline models that rely on distributional semantics and geometric distances. The result from our experimental evaluation clearly show that our proposed approach outperform the existing baseline models.

The remainder of this paper is organized as follows: Section 2 briefly discusses about the state of the art. Section 3 presents our proposed information theoretic measure of novelty detection along with the two baseline techniques that are further used to evaluate our proposed model. Section 4 presents the experimental set-up, description of the dataset and the results obtained. Finally, in section 5 we conclude this paper.

2. State of the Art

A recent work by (Verheij, Kleijn, Frasincar, & Hogenboom, 2012) presents a comparative study of different novelty detection methods evaluated on news. Novelty detection task was also presented in TREC 2002-2003 (Harman, 2002) (Voorhees, 2003). Here, novelty detection was examined at sentence level (Allan, Wade, & Bolivar, 2003) (Li & Croft, 2005) (Kwee, Tsai, & Tang, 2009) (Tsai, 2010). Later on in TREC 2004, 13 participants have proposed different features and algorithms to perform the desired task (Gamon, 2006). Some of the key features include string based comparisons, synonymy resolution, co-reference resolution and named entity recognition (e.g. (Blott, et al., 2004) (Zhang, Xu, Bai, Wang, & Cheng, 2004) (Abdul-Jaleel, et al., 2004) (Eichmann, et al., 2004). Thresholds are either determined based on a notion of mean score, or are determined in an ad hoc manner. Alternatively, Tomiyama et al (Tomiyama, et al., 2004)(2004), use an SVM classifier to make the binary classification of a sentence as novel or not. Blott et al. (2004) uses a tf.idf based metric of "importance value" at an ad hoc threshold. Tomiyama et al. (2004) using an SVM classifier trained on 2003 data, they have used features based on conceptual fuzzy sets derived from a background corpus. Abdul-Jaleel et al. (2004) used named entity recognition along with cosine similarity as a metric. (Schiffman & McKeown, 2004) used a combination of tests based on weights for previously unseen words with parameters trained on the 2003 data set. Allan et al. (Allan, Wade, & Bolivar, 2003) evaluated seven measures for novelty detection separating them in word count measures and language model measures. The results showed that the maximum cosine similarity between a sentence and a number of previously seen ones, works as well as complex language model measures. (Schiffman & McKeown, 2004) proposed a linear combination of the maximum cosine similarity measure with a metric that aggregates the TF-IDF scores of the terms in a sentence. Similar approach was taken by (Margarita, Rousseau, Ntoulas, & Vazirgiannis, 2014) to identify document level novelty detection of document stream. A Vector Space model based technique is also proposed by (Zhao, Zhang, & Ma, 2006). A more sophisticated metrics are defined on the basis of graph representations (Gamon, 2006). Here, the content of a text document is represented in the form of a graph with the terms as a vertices and their relationship as edges.

3. The Proposed Novelty Scoring Method

We have considered that each document *d* is represented by a bag-of-word as, $< (q_1, w_1), (q_2, w_2), \dots, (q_n, w_n) >$ where q_i is the i[th] unique term in document *d* and w_i is the corresponding weight computed with respect to a collection of documents *C*. The novelty score of each document is computed with respect to the previously stored documents (represented by C) in the system. Therefore, for each new document d, a novelty score NS(d, C) is computed, indicating the novelty of this document for the given document collection. In the described context, we have declared a document d_i as *novel* when the corresponding novelty score NS(d_i, C) is higher than a given threshold θ.

In this paper, we have introduce an information theoretic approach towards determining the novelty score of a given document. We have defined the novelty of a document in terms of its *information content*. Thus, higher the information content of a document is higher is the chance of it being novel. The information content is a heuristic measure for term specificity and is a function of term use. More generally, by aggregating all the information content of the terms of a document, it can be seen as a function of the vocabulary use at the document level. Hence, our idea to use it as an estimator of novelty – a novel document being more likely to use unique vocabulary than the ones in the other documents. In a way, a document is novel if its terms are also novel – i.e. previously unseen. This implies that the terms of a novel document have a generally high information content. We have computed the information content of a document in terms of its *Entropy*. We define the entropy of a text T, with λ words and *n* different ones as:

$$E_T(p_1, \dots, p_n) = \frac{1}{\lambda} * \sum_{i=1}^{n} p_i \, (\log_{10} \lambda - \log_{10} p_i)$$

Where, $p_i (i = 1 \dots n)$ is the probabilistic measure of the specificity of the i^{th} word in the text T. The technique to compute the term specificity is discussed in the subsection below. In order to avoid the problem of zero probabilities we have used linear interpolation smoothing, where document weights are smoothed against the set of the documents in the corpus. Then the probabilities are defined as:

$$\theta_{d_n}(q) = \lambda * \theta_d(q) + (1 - \lambda) * \theta_{d_1 \dots} \theta_{d_{n-1}}(q)$$

where, λ ∈ [0, 1] is the smoothing parameter and is the probability of term q in the corpus C. In our experiments, λ was set to 0.9.

Apart from computing the information content of the whole document, we have also focused on determining the importance of the individual terms in determining the novelty of a document. As discussed earlier, the core of our novelty prediction engine is to compute the rarity of a document which can in turn be computed by determining the rarity of the individual terms. Accordingly, we have applied the principle of *Inverse Document Frequency* (IDF) as discussed in (Margarita, Rousseau, Ntoulas, & Vazirgiannis, 2014). It has been established that IDF incorporates a heuristic measure that determines a term's specificity and thus, is a function of the term's usage. Therefore, aggregating all the IDF of the terms of a given document may led us to a better estimator of the documents novelty. IDF is originally defined as, $IDF(q, C) = \log\frac{N}{df_q}$ where, q is the term in hand, df_q is the document frequency of the term q across the corpus C and N is the total number of documents in the collection. On the other hand, in probabilistic terms IDF can be computed as: $IDF_p(q, C) = \log\left(\frac{N-df_q}{df_q}\right)$. It is to be noted here that this IDF definition can lead us to negative values if a term q appears in more than half of the documents. This property could be of importance as we want to penalize the use of terms appearing in previously seen documents. In extreme cases where the document frequency of a term q is null or equal to the size of the corpus, we have incorporated a smoothing variant where the term frequency is usually added by 0.5 to both numerator and denominator).

Baseline Approaches

Cosine Similarity Model based Approach: We have used the vector space methods based on cosine similarity as a baseline approach for our novelty detection purpose. It assumes if two or more documents are very similar to each other, the information they contain were already seen before and cannot be considered as novel. We also introduce a second baseline using the *Mean and Max Cosine Similarity*. Where, a document is novel if its mean/max similarity to the documents in the corpus is below/above a threshold. The mean/max similarity for two documents are

$$CS_{Max}(d_1, C) = \max_{1 \le i \le |C|} CS(d_1, d_2) \text{ and}$$

$$CS_{Mean}(d_1, C) = \frac{\sum_{i=1}^{|C|} CS(d_1, d_i)}{|C|}$$

KL divergence based Approach: We have used minimum Kullback-Leibler (KL) divergence as another baseline approach for our novelty detection task (Verheij, Kleijn, Frasincar, & Hogenboom, 2012). Thus, the respective novelty scoring formula is as follows: $MinKL(d_1, C) = \min_{1 \le i \le |C|} KL(\theta_{d_1}, \theta_{d_2})$.

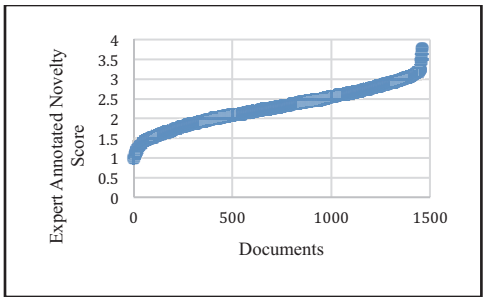

Fig. 1: Distribution of expert annotated avg. novelty score.

4. Evaluation

The Dataset: We have collected around 1600 ideas from a real world innovation contest. Here, participants are asked to express ideas relevant to a specific organization. Length of the text documents varies from 10 to 6000 words. Average number of words per document is 400. Each of the text documents were manually evaluated by a group of six experts. The experts were asked to grade the documents according to their novelty. The novelty score ranges between 0 and 5. 0 represents poor novelty and 5 represents high novelty value.

We have used the Fleiss Kappa statistics to compute the inter annotator agreement of a partial sub-set of data for which detailed manual scores for each of the expert annotators were available. We found Kappa, $\kappa = 0.4$ which reflects a marginal agreement between the experts. Finally, we have computed the average of the expert annotated novelty score for each of the documents. We have encountered around 100 documents containing either no text materials or contain texts written in non-English languages. Thus, these 100 documents were discarded from our further analysis. Figure 1 represents the distribution of the average novelty score of the final set of 1500 text documents.

Based on the final novelty score, we have broadly classified each of the documents into three different classes namely, *high novelty, average novelty* and *low novelty class*. The *high novelty* class represent documents that contains some unique concepts with respect to the other existing documents. On the other hand, *low novelty* documents represent concepts that are very common and talked by most of the writers. The classification of the documents are done by computing the *Average* and *standard deviation* of each of the expert annotated novelty scores. Therefore, a document is *highly novel* if its expert annotated novelty score is above the $average + stdev$ threshold; similarly, a document is *less novel* if its expert novelty score falls below the $average - stdev$ threshold and the rest of the documents lies under the *average novelty* class.

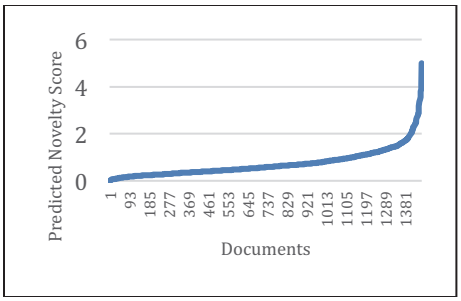

Fig 2: Novelty score predicted by the proposed model

	P	R	F
MaxCosine Similarity	58.4%	68.5%	62.7%
MinCosine Similarity	55.3%	57.1%	56.1%
KL-Divergence	66.3%	71.7%	68.8%
New Novelty Score	74.0%	77.0%	75.4%

Table 1: Performance of the different models

Type of noise	**Examples**
Unhandled html tags	* *, &amp
Non-English terms	BİZ, BİR, AİLEYİZTüm personelin, Amaç:Böylece
Spelling errors	Informtion, beeing, becouse
Concatenated words due to incorrect import/export from system	Studentjob, giftcertificates, Purpose:To, Example:My, service.We, DIRECTClient, Bank.If
Multiple concatenated words	Companyproducesorownerreachesretirement, capitalseekingprofitableinvestment

Table 2: Analysis of different type of noisy words in documents

Evaluation Method and Metrics: Each of the 1500 documents were passed to the two baseline approaches (*Cosine Similarity* model and *Language Model*) as well as our proposed *Information Theoretic* model. Each of these three model returns a novelty score of a document with respect to the rest of the document collection. The distribution of the predicted novelty score of each document across each of the baseline and proposed model is depicted in Figure 2. In the x-axis we have the documents and y-axis plots the respective novelty score of the documents.

The novelty score of each documents are then compared with the expert annotated average novelty score values. We have also computed the Spearman's correlation between the computed novelty score and the expert annotated novelty score. Next, we have classified the predicted novelty score values into the three different classes following the same technique discussed for the expert annotated novelty score. Therefore, documents with novelty score above $average + stdev$ are predicted to be in the *high novelty* class, novelty score in the range of $average - stdev$ belongs to *low novelty* class, and the rest belongs to *average novelty class.* Finally, we have compared the automatic classification with that of the expert annotated manually classified dataset. The performance of the Novelty Detection models are evaluated in terms of the three standard parameters: Precision (P), Recall (R) and F-Measure (F).

Results and Analysis

As discussed earlier, we have used the expert annotated dataset to evaluate the performance of our proposed method and the three baseline methods that are already in use for this task. All the above methods takes into account the similarity or divergence among the document in hand and the existing collection of documents C and rate it as *Novel* based on a pre-computed threshold value. The performance of the models in terms of precision, recall and F-measure are tabulated in Table 1. Here, we can observe that the language model based KL-divergence approach performs slightly better (F-measure 68.89%) than both the variation, MaxCosine (F-measure 62.79) and MinCosine (61.11) of the distributional semantics based approach to novelty detection. On the other hand, the proposed entropy based information theoretic approach achieves a much higher F-measure of 75.4%. Thus, it is clear that the performance of our proposed technique well surpasses the performance of the existing baseline approaches.

In general, we have observed around 23% of the input text documents were miss-classified. We categorize the error types into five cases. In Case-I, There are around 32% text documents found where the same texts are copied repeatedly over multiple times. This is unnecessarily making the length of the document high. As the computed novelty score is a function of the document length and the information content of the individual words, repetition of words is significantly affecting the model's performance. Case-II contains 32% document sets that are too short in word length. These documents are approximately 10-20 word long. Consequently, it becomes difficult to extract any informative terms or phrases that can contribute to the information content or entropy of the documents. Therefore, in all such cases our proposed model predicted the input document to be under *low novelty* class. In Case-III, we have encountered around 5% documents which contains a substantial number of foreign language words mixed with English. Since, these words rarely occurred in the corpus, they significantly increasing the entropy score of a document which in turn get biased towards high novelty score. In case-IV, we found a small set of documents (1%) that referred some urls and images that contributes to the expert annotated novelty score. However, extracting information from such heterogeneous sources is not the scope of the current work. Thus, our model fails to identify such cases. Finally, in Class-V we encountered 30% text documents containing

too many noisy words which significantly affects the overall novelty score of a document. Table 2 enumerates different types of noises in the text document that results in the deviation of predicted novelty score with the expert annotated score.

5. Conclusion

In this paper we have proposed a new information theory based method to automatically determine the novelty score of a given document with respect to the other existing ones. We have considered an idea to be novel if its information content exceeds a certain threshold. The information content of a document is computed in terms of its entropy. We have also computed the specificity of each of the individual terms in a document using the probabilistic Inverse Document Frequency score. We have compared the performance of our proposed model with the expert annotated data along with some of the commonly used baseline models that are discussed in this paper. The result from our experimental evaluation clearly show that our proposed approach outperforms the existing baseline models. We have observed that around 23% of the cases our model fails to correctly classify the documents. This may be due to a number of reasons discussed in this paper. In the next phase of our work we intend to focus on those aspects that led to the miss-classification.

References

Abdul-Jaleel, N., Allan, J., Croft, W. B., Diaz, F., Larkey, L., Li, X., . . . Wade, C. (2004). *UMass at TREC 2004: Novelty and HARD.* DTIC Document.

Allan, J., Wade, C., & Bolivar, A. (2003). Retrieval and novelty detection at the sentence level. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 314-321). ACM.

Blott, S., Boydell, O., Camous, F., Ferguson, P., Gaughan, G., Gurrin, C., . . . Smeaton, A. F. (2004). *Experiments in terabyte searching, genomic retrieval and novelty detection for TREC 2004.* NIST.

Chesbrough, H. W. (2006). *Open innovation: The new imperative for creating and profiting from technology.* Harvard Business Press.

Eichmann, D., Zhang, Y., Bradshaw, S., Qiu, X. Y., Zhou, L., Srinivasan, P., . . . Wong, H. (2004). *Novelty, Question Answering and Genomics: The University of Iowa Response.* TREC.

Gamon, M. (2006). Graph-based text representation for novelty detection. *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing* (pp. 17-24). Association for Computational Linguistics.

Harman, D. (2002). Overview of the TREC 2002 Novelty Track. *TREC.*

Kwee, A. T., Tsai, F. S., & Tang, W. (2009). Sentence-level novelty detection in English and Malay. In *Advances in Knowledge Discovery and Data Mining* (pp. 40-51). Springer.

Li, X., & Croft, W. B. (2005). Novelty detection based on sentence level patterns. *Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 744-751). ACM.

Lin, Y.-l., & Brusilovsky, P. (2011). Towards open corpus adaptive hypermedia: a study of novelty detection approaches. In *User Modeling, Adaption and Personalization* (pp. 353-358). Springer.

Margarita, K., Rousseau, F., Ntoulas, A., & Vazirgiannis, M. (2014). Using temporal IDF for efficient novelty detection in text streams. arXiv preprint arXiv:1401.1456.

Markou, M., & Singh, S. (2003). Novelty detection: a reviewâ€"part 1: statistical approaches. *Signal processing*, 2481-2497.

Schiffman, B., & McKeown, K. (2004). *Columbia University in the Novelty Track at TREC 2004.* TREC.

Tomiyama, T., Karoji, K., Kondo, T., Kakuta, Y., Takagi, T., Aizawa, A., & Kanazawa, T. (2004). *Meiji University Web, Novelty and Genomic Track Experiments.* TREC.

Tsai, F. S. (2010). Review of techniques for intelligent novelty mining. *Information Technology Journal*, 1255-1261.

Tsai, F. S., & Kwee, A. T. (2011). Experiments in term weighting for novelty mining. *Expert Systems with Applications*, 0957-4174.

Verheij, A., Kleijn, A., Frasincar, F., & Hogenboom, F. (2012). A comparison study for novelty control mechanisms applied to web news stories. *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on* (pp. 431-436). IEEE.

Voorhees, E. M. (2003). *Overview of TREC 2003.* TREC.

Zhang, H., Xu, H., Bai, S., Wang, B., & Cheng, X. (2004). *Experiments in TREC 2004 Novelty Track at CAS-ICT.* TREC.

Zhao, L., Zhang, M., & Ma, S. (2006). The nature of novelty detection. *Information Retrieval*, 521-541.