Prototype/Topic Based Clustering Method for Weblogs

Fernando Perez-Tellez[a,*], John Cardiff[a], Paolo Rosso[b], David Pinto[c]

[a] *Social Media Research Group, Institute of Technology Tallaght, Dublin, Ireland*

[b] *NLE Lab. – PRHLT Research Center, Universitat Politècnica de València, Spain*

[c] *FCC, Benemérita Universidad Autónoma de Puebla, Mexico*

Abstract. In the last 10 years, the information generated on weblog sites has increased exponentially, resulting in a clear need for intelligent approaches to analyse and organise this massive amount of information. In this work, we present a methodology to cluster weblog posts according to the topics discussed therein, which we derive by text analysis. We have called the methodology *Prototype/Topic Based Clustering*, an approach which is based on a generative probabilistic model in conjunction with a Self-Term Expansion methodology. The usage of the Self-Term Expansion methodology is to improve the representation of the data and the generative probabilistic model is employed to identify relevant topics discussed in the weblogs. We have modified the generative probabilistic model in order to exploit predefined initialisations of the model and have performed our experiments in narrow and wide domain subsets. The results of our approach have demonstrated a considerable improvement over the pre-defined baseline and alternative state of the art approaches, achieving an improvement of up to 20% in many cases. The experiments were performed on both narrow and wide domain datasets, with the latter showing better improvement. However in both cases, our results outperformed the baseline and state of the art algorithms.

Keywords. Short Text Analysis, Weblog Clustering, Topic Identification.

* Corresponding author: Fernando Perez Tellez, SMRG – Institute of Technology Tallaght Dublin, Tallaght, Dublin 24, Ireland.
E-mail: fernandopt@gmail.com

1 Introduction

The Internet has witnessed changes on a huge scale in recent years. It has become a new tool of interaction and socialisation among Internet users, all of which is part of the evolution of the WWW towards the "social web", i.e., Web2.0 applications such as wikis, weblogs, and social networks. The improvements in computing technology and in connection speeds have made the new web more accessible for everyone.

An important part of the social web is the blogosphere. It is a decentralised medium of expression and interaction for everybody that makes it possible to share ideas and spread opinions. Nowadays we can find weblogs on almost any subject. They are usually considered as “short text”. From the statistical perspective [39] a short text is described as text that does not have enough content from which a meaningful statistical model can be built. The average length of a weblog can vary across different weblog sites, whereby the main post can contain between a couple of sentences and 500 words but the postings (comments or feedback) from other users can be very short, consisting of one or two sentences.

In order to deal with the huge amount of information published every day on the blogosphere, there is a clear necessity for intelligent systems and applications that can manage and provide an automatic analysis and organisation of this kind short text documents, with the objective of providing efficient manipulation of the information and retrieving efficacious information required for the user.

The principal approaches for the automatic organisation of documents are based on common classification or clustering methods [2] [3]. In classification is not easy to have training subsets for particular domains. On the other hand, the time required to compute the similarity measures among documents is often prohibitive for clustering approaches.

The motivation of this research work is to implement a methodology which can organise automatically weblog posts into topic based clusters, so that the manipulation and information retrieval can be performed more accurately. Our approach is based on the assumption that there is little or no information that can be exploited by a classification approach. For this reason, we consider that document clustering – the assignment of documents to previously unknown categories – is an appropriate solution to the purpose of categorising weblogs [41], rather than classification. The latter approach would require providing tags of categories in advance, but in real scenarios we usually deal with information from the blogosphere without knowing the correct category tag or at least with very limited information about their categories.

As stated in [29] and [11], weblog posts can usually be characterised as short texts and with a general writing style. These are undesirable characteristics from a clustering perspective, as typically insufficient discriminative information is provided. In order to improve these particular characteristics of weblogs, we employ an enrichment method named the Self-Term Expansion Methodology [32] that does not use external resources, relying only on information included in the corpus itself. We demonstrate that the application of this methodology can improve the quality of topic clusters, and further that the improvement will be more significant where the corpus is composed of well-delimited categories which share a low percentage of vocabulary (i.e., a wide domain corpus).

This paper describes an approach for clustering data, specifically weblog posts, according to their topic of discussion. We are particularly interested in ways of guiding the clustering process, and for this purpose, we have employed a topic detection method [9]. The value of this kind of method is the strong theoretical framework with the idea that each document is a mixture of topics, where topics are distributions over words.

Topic detection and tracking is a well-studied area [4] [5], which focuses on extraction of significant topics and events from documents (such as news articles). In our case we are using topic detection to cluster weblog data. We introduce and evaluate a novel methodology for clustering weblog posts called Prototype/Topic Based Clustering which is based on a topic detection method that is used in the identification of latent topics over text This is complemented with an expansion methodology in order to improve the document representation, thereby increasing the discriminative information of the topics discussed in weblogs.

The rest of this paper is organised as follows. Section 2 presents the related work. Section 3 explains the proposed approach and the techniques used in this research work. Section 4 describes the dataset used, the experiments, the obtained results and a comparison with a baseline algorithms. Section 5 provides an analysis of results. Finally, in Section 6 we present the conclusions.

2 Related Work

We consider the topic detection task as the problem of finding the most prominent topics in a collection of documents; in general terms, identifying a set of words that constitute topics in a collection of documents. There are previous attempts at topic detection in online documents such as in [15], where the authors present a topic detection system composed of three modules that attempt to model events and reportage in news. The task of finding a set of topics in a collection of documents has also been attempted in [45], in which the authors based their approach on the identification of clusters in keywords that are taken as representation of topics. They have employed the well-known *k*-means algorithm to test some distance measures based on a distribution of words.

Topic detection is also addressed in [38], where the authors present a method which uses bloggers' interests in order to extract topic words from weblogs. In this approach the authors assume that topic words are words commonly used by bloggers who share the same interests, and they use these topic words to compute similar interests between each two bloggers by using the cosine similarity measure.

Topic detection has also been applied to research papers. In [37] the authors cluster them into hierarchical overlapping clusters using the topics discussed in them as a similarity measure. The authors ranked the research papers in topic clusters by using a modified Page-Rank algorithm. This approach was developed and focused on a very narrow domain i.e., research papers documents in the computing domain.

The clustering of weblogs has become an active topic of research. For instance, in [25] the authors build a word-page matrix by downloading weblog pages and apply the *k*-means clustering algorithm with different weights assigned to the title, body, and comment parts. In [1], the authors use weblog categories to build a category relation graph in order to join different weblog classes; they use edges in the category relation graph to represent similarity between different categories and they represent nodes as categories.

Another approach which uses topic detection methods in weblogs is presented in [44] in which the authors describe a topic detection approach based on n-grams (a subsequence of items from a given sequence usually words or letters). A research work which uses topic detection for clustering microblogs is described in [48]. The approach augments lexical evidence for topical similarity using Wikipedia[1] as an external resource. The idea is to relate microblog posts to Wikipedia pages therefore semantic similarity can be estimated.

[1] http://www.wikipedia.org/

An approach for clustering short messages called the Multi-stage Clustering algorithm is presented in [42], where the authors focused on clustering tweets using a clustering framework that is broken into two distinctive tasks. The first task is batch clustering of user annotated data, which allows the conversion of document clustering task to a tag clustering problem. The second task is the online clustering of a stream of tweets which uses the centroids generated in the previous stage in order to assign each new message to a cluster. The tag clustering is done in batch mode and the actual tweet clustering is done in an online manner.

In terms of topic detection models there are probabilistic models that have been proposed and are based upon the idea that documents are mixtures of topics and a topic is a probability distribution over words. In [18], a probabilistic approach to semantic representation is presented which models the probability with which words occur in different contexts, capturing the relationships between words. In [17], a generative model for documents is introduced which is used to identify the content of a document. The authors present a Markov chain Monte Carlo algorithm for inference in this model. A probabilistic latent semantic analysis (PLSA) model has been proposed [20] in order to analyse and extract latent topics in text documents. There are some variations such as in [49] that focus the extracted topic models on the content words rather than on the usual words in the collection. The research work of [23] presents a methodology to cluster legal documents based on the topics discussed therein.

Our approach is focused on detecting the topic clusters contained in the corpus itself. The novel aspect is based on using a topic detection method (guided or not) to identify possible references that could be used in the clustering process, and the expansion methodology in order to improve the representation of the weblogs. Our approach is domain and language

independent and does not require any external linguistic resource to be used. In addition, the clustering process is a simple well-known method that demonstrates fast performance compared to hierarchical clustering algorithms such as *k*-means.

3 Prototype/Topic Based Clustering Methodology

In this section, we present the methodology we used in order to improve the quality of clusters and which clusters weblog posts using prototypes as references. A prototype is composed of keywords of a topic discussed over the weblog posts, i.e., words which identify a topic discussed in the weblog corpus. We refer to our approach as Prototype/Topic Based Clustering (P/TB Clustering). Our approach is based on a topic detection method that produces prototypes (vectors of keywords) that are used as reference points in the clustering process. An initial version of the methodology was presented in [33] and in this work we have improved and extended it as follows: new baselines for comparison purposes have been defined. A new comparison of results of our clustering method against a standard clustering algorithm. Another important improvement is the introduction of two new forms of initialisation in the prototype construction based on the topic identification method.

Our approach is composed of an expansion procedure which is an adaptation of the Self-Term Expansion Methodology (S-TEM) [32], which is followed by the application of the Latent Dirichlet Allocation model (LDA) [9] that feeds into the prototype/topic based clustering process.

The steps of the methodology are:

- *Self-Term Expansion Methodology (S-TEM).* This is an expansion methodology whose purpose is to improve the characteristics of the text from a clustering perspective. It consists of the following techniques:

- Self-Term Enriching Technique. This step improves the representation of short documents by using a term enriching procedure. We use only the information being clustered to perform the term expansion, i.e., no external resource is employed, as it is often difficult to identify appropriate linguistic resources for information such weblogs.
- Term Selection Technique. This is applied in order to select the most important and discriminative information of each category, thereby reducing processing time for the subsequent stages of our approach (topic detection and clustering).

- *Topic Identification*. This task is used to identify the relevant topics discussed in the collection, so the topics detected are used to create references of the categories. In other words, we apply a Topic Identification approach to detect the latent topics over the weblogs posts and with these topics construct the prototypes (one for each category) which will be used in the clustering process. The topic detection model that we have employed is the Latent Dirichlet Allocation (LDA) method [9], which is a generative probabilistic model for discrete data.

  The step of selecting the topics can be initiated automatically without any input, in other words the initial parameters are automatically estimated. Alternatively, the process can be initially guided by keywords (usually nouns) that occur frequently in each category. The process of initialisation of the Topic Identification process can be manual or semi-automatic by selecting relevant words per category. In this research work, we present three variations of the Topic Identification method (unguided initialisation, manually guided initialisation and semi-automatic guided initialisation) for comparison purposes. The output of this stage is a set of prototypes which

represent the categories in the collection, i.e., they are lists of relevant words discussed in the weblog documents.

- *Clustering process.* The prototypes constructed in the previous step are used in this process which will create clusters by comparing each weblog post to each prototype. A weblog post is assigned to the cluster for which it obtains the highest similarity value to the corresponding prototype. The Jaccard coefficient measure [27] as used as the similarity measure to form the clusters.

Figure 5 shows the complete process of the prototype/topic based clustering approach. In the following sub-sections we describe in detail each of the steps of our methodology. The motivation of using an expansion technique is to improve the representation of weblog text in order to highlight the relevant information which is used to obtain better clustering results. In addition, we use the intermediate step of prototype generation in order to create them from the main topics of the whole set of weblogs. In other words, the intermediate step is needed due to the fact that we are dealing with short text and the probability of including the main topics in all the small documents is low. For this reason, we consider it appropriate to construct the groups from the most probable topics from the whole collection rather than topics from single documents. The consequences of not using the prototypes are that the topics of some documents that may not be relevant for the whole collection and it may produce multiple clusters with few elements.

3.1 Step 1: Self-Term Expansion

The Self-Term Expansion Methodology (S-TEM) comprises a twofold process: the Self-Term Enriching Technique, which is a process of replacing terms with a set of co-related terms, and a Term Selection Technique with the role of identifying the relevant features.

The idea behind Term Expansion has been studied in previous works such as [34] and [16] in which external resources have been employed to determine the correct sense of a word given in context. Term expansion has been used in many areas of natural language processing such as word disambiguation in [7], in which WordNet [14] is used in order to expand all the senses of a word. There are proposals of using dictionaries such as [24] and [46] widely used in word sense disambiguation. In [21] and [28] different ways of improving text clustering by employing ontologies, authors have reported the improvement of the similarity intra-documents by incorporating background knowledge from external resources such as WordNet.

While enrichment of terms using an external knowledge source is valuable, the application of term expansion by using co-related terms will only improve the baseline results if we carefully select the external resource to use. In other words, we would need to know the domain of the documents to be cluster a priori. In addition, for particular domains it may not be possible to identify an external resource. Therefore, we consider the use of an automatic and domain independent constructed lexical resource to be the best option. There are some proposals such as [35] and [33] where words are expanded with co-occurrence words for word sense disambiguation. However, in the particular case of the S-TEM methodology no external resource is employed and it has shown good improvement in the representation of the documents, in particular for the clustering task.

The technique consists of replacing terms of a weblog post with a set of co-related terms. We consider it particularly important to use the intrinsic information of the dataset itself as it is difficult to identify an appropriate external resource due to the rapidly changing content of weblog posts. A co-related list is calculated from the target dataset by applying the Pointwise Mutual Information (PMI) [27]. PMI provides a value of relationship between two words;

however, the level of this relationship must be empirically adjusted for each type of text. In this work, we empirically established a value greater than 2 to be the best threshold. In other experiments we have conducted using more formal texts [32], a threshold of 6 was used; however, in weblog documents co-related terms are rarely found. This set of co-related terms will be used to expand every term of the original corpus. The appropriate value of each of the other parameters used in the expansion process was established using a range of different datasets in other experiments. They have demonstrated behaviour consistent with results reported in previous experiments [30] [32].

The Self-Term Enriching Technique is defined formally as follows: Let $D = \{d_1, d_2, \ldots, d_n\}$ be a document collection with vocabulary $V(D)$. Let us consider a subset of $V(D) \times V(D)$ of co-related terms as $RT = \{(t_i, t_j) | t_i, t_j \in V(D)\}$ The RT expansion of D is $D' = \{d'_1, d'_2, \ldots, d'_n\}$, such that for all $d_i \in D$, it satisfies two properties: 1) if $t_j \in d_i$ then $t_j \in d'_i$, and 2) if $t_j \in d_i$ then $t'_j \in d'_i$, with $(t_j, t'_j) \in RT$. If RT is calculated by using the same target dataset, then we say that D' is the Self-Term Expansion version of D. The degree of co-relation between a pair of terms is determined by a co-related method, which is based on the assumption that two words are semantically similar if they occur in similar contexts [19].

The Term Selection Technique helps us to identify the best features for the clustering process. However, it is also useful to reduce the computing time of the clustering algorithms. In particular, we have used Document Frequency (DF) [40], which assigns the value DF(t) to each term t, where DF(t) is the number of posts in a collection in which t occurs. The Document Frequency technique assumes that low frequency terms will rarely appear in other documents; therefore, they will not have significance on the prediction of the class of a document.

In more detail, the enriching process is performed by constructing a co-relation model among all the words using PMI, i.e., calculating PMI for each pair of words. Then a co-occurrence list is generated by filtering some relationships applying some thresholds such as PMI greater than 2 and word frequency greater than 3. After the enriching process we have used the Term Selection Technique then we have selected from 10% to 90% of vocabulary of the expanded text, in order to confirm the minimum percentage of vocabulary which can provides the best input to the Topic Identification process.

3.2 Step 2: Topic Identification

In general, a topic model is a hierarchical Bayesian model that assigns to each document a probability distribution over topics. We have adapted the Latent Dirichlet Allocation (LDA) model [9] which is derived from the idea of discovering short descriptions of the members of a collection, in particular discrete data, in order to allow efficient processing of huge collections, while keeping the essential statistical relationships that may be used in other tasks such as classification.

The LDA model is based on an assumption that the words of each document arise from a mixture of topics, each of which is a distribution over the vocabulary. Documents that discuss similar topics will use similar group of words. LDA tries to detect groups of words which frequently occur in a collection of documents. This method has been used for automatically extracting the topical structure of large document collections. In other words, it is a generative probabilistic model of a corpus that uses different distributions over a vocabulary in order to describe the document collection.

The use of topic models has been an area of considerable interest for pattern recognition researchers. LDA in particular has become very popular and effectively applied to text-related tasks. It has been used in several applications such as entity resolution [8], fraud

detection in telecommunication systems [47], image processing [13], topic detection in text [17], reputation management [6], and word sense disambiguation [12]. In [9] the authors describe the model as a generative probabilistic model for discovering latent semantic topics in large collections of data.

LDA relies on the co-occurrence of the words in documents to assign the documents to certain topics. The original LDA model is a completely unsupervised approach which models documents as a mixture of topics. This model produces automatic summaries of topics in terms of a discrete probability distribution over words for each topic and infers discrete distributions per-document over topics.

The approach is similar to *probabilistic Latent Semantic Index* (pLSI) [20], in which the main idea is to model each word in a document as a sample from a mixture model, in which the components of the mixture are multinomial random variables that can be viewed as words generated from topics. However, LDA may be seen as a step forward with respect to pLSI as it provides no probabilistic model at the level of documents [9].

The Topic Identification method allows defining the number of keywords to be extracted, in this sense we have varied the number of keywords selected from 100 to 3,000 in order to confirm the best and minimum number of terms for the clustering task. This step generates a set of prototypes, one per category, each containing a list of keywords. The prototypes were constructed with terms with the highest probability in each topic of the whole collection. These prototypes form the input to the final step which is the clustering process.

3.3 Step 3: Clustering Phase

The input to this phase process is the set of prototypes (each of which comprises a list of keywords corresponding to a specific topic) and the original weblog posts. The task is to assign each post to a cluster according to the most similar prototype.

We have chosen a clustering approach based on the Jaccard coefficient as the similarity measure for reasons of efficiency. Its low computational resources allow the approach to be easily scaled. The outputs of this phase are clusters; they are created based on the prototypes given as inputs.

4 Experiments

In this section, we present the datasets, the experiments and results obtained using the approach presented in this research work. We also define baselines by which we can establish the improvement obtained by using our approach. Firstly, we improved the representation of the posts by applying the S-TEM methodology so that the Topic Identification process can build better prototypes. We empirically found that the enrichment process gave optimal results when 10% of the most relevant vocabulary was selected (i.e., terms with the highest co-relation value). In other words, co-related terms were evaluated and we selected 10% of the vocabulary, having the highest co-relation value. Subsequently, the corpus was enriched with this vocabulary.

In the Topic Identification phase, we performed experiments with different initialisations of the Topic Identification method as well as an unguided version. Initialisation is the process of providing a predefined list of initial topics to the Topic Identification process. We have used two different approaches for select the initial topics. The first approach is the manual selection of words which would be potentially important for the categories. In the second, we have used a semi-automatic method of selecting the relevant words. The method we applied is the Transition Point (TP) Technique described in [31]. This technique produces a frequency value list which orders the vocabulary according to each term's relative frequency in the document and then splits the vocabulary of a document into two sets of terms, low and high frequency to identify what is called Transition Point (middle point). It is based on the

Zipf's law of word occurrences [50] and the hypothesis behind it is that the medium frequency terms are closely related to the conceptual content of documents.

Our approach uses unweighted set in the construction of the prototypes because our intension is to keep the approach as simple as possible for the same reason Jaccard's coefficient was employed as a similarity measure in the clustering process and our results have shown improvement over the baselines. We also would like to mention that we have used the original inference algorithm in the estimation of α and β parameters for LDA method proposed by the authors [9]. It is a generative probabilistic model that uses inference techniques based on variational methods and an expectation maximization algorithm for empirical Bayes parameter estimation.

The structure of this section is as follows: firstly we describe the dataset used in our experiments. Secondly, we explain the construction of the baselines against which we benchmark the obtained results. We then present the various experiments we conducted and the results obtained.

4.1 Description of the Datasets

In this section, we describe the datasets used in our experiments. We have constructed two datasets, both of which are subsets of the ICWSM 2009 Spinn3r Blog Dataset[2]. The data is in XML format and according to the Spinn3r crawling[3] documentation, it is further arranged into tiers, approximating search engine ranking to some degree. The weblog posts are treated as raw text, i.e., we have not used any additional information provided by the XML tags. As a pre-processing step, we have removed stop words – high-frequency words that have no significant meaning in a phrase – and punctuation symbols.

[2] The corpus was initially made available for the 2009 Data Challenge at the 3rd International AAAI Conference on Weblogs and Social Media, http://www.icwsm.org/2009/data/

[3] http://spinn3r.com/documentation/

We have focused the experiments carried out on the "Yahoo Answers", weblog site[4] – in which people share what they know and ask questions on any topic of interest to the user, in order to be answered by other users. We have extracted from this corpus two distinct subsets (see Table 1). The category name in this table was taken from the category tag provided in the collection.

The first subset contains 10 categories with 25,596 posts and vocabulary size of 66,729. It may be considered as "narrow domain", because the vocabulary in the categories is quite similar. As we have seen the categories in this subset are more difficult to distinguish, being predominantly technology related. The second subset contains 10 categories with 48,477 posts and a vocabulary size of 122,960 terms.

As opposed to the narrow domain subset, it may be considered to be "wide domain" because its categories have a low overlapping degree of vocabulary due to the fact that the topics discussed in this subset are different and shared terms among categories is low.

The process of clustering narrow domains brings additional challenges because the categories in the collection share common terms. Moreover, the shortness of this kind of data makes this task even more challenging. For this reason, we expect to have better clustering results when dealing with wide domain than with narrow domain.

The purpose of constructing two subsets with these characteristics is to demonstrate the effectiveness of the method herein across both wide and narrow domains, and also to test the relative effectiveness of the approach in each case.

[4] http://answers.yahoo.com/

The category tags given in the collection were used for gold standard construction purposes. They are shown in Table 1 to provide a better idea of the subsets used in the experiments.

4.2 Evaluation Measure Definition

We have used the well-known *F-measure* to evaluate our experiments which is composed of *precision* and *recall* metrics which are well-known measures used in evaluating the effectiveness of a system.

The F-Measure [43] is defined as follows: given a set of clusters $C=\{C_1,\ldots,C_{|C|}\}$ and a set of classes $C^*=\{C_1^*,\ldots,C^*_{|C^*|}\}$, the F-Measure between a cluster C_i and a class C_j^* is given in the following equation.

$$FMeasure(C_i, C_j^*) = \frac{2 * precision(C_i, C_j^*) * recall(C_i, C_j^*)}{precision(C_i, C_j^*) + recall(C_i, C_j^*)}$$

The global performance of a clustering method is computed using F-Measure values, the cardinality of the set of clusters obtained, and normalising by the total number of documents |D| in the collection. The obtained result is the F-Measure and it is shown in the next equation.

$$F-Measure = \sum_{1 \le i \le} \frac{|C_i|}{|D|} \max_{1 \le j \le |C^*|} FMeasure(C_i, C_j^*) \quad \text{where } 1 \le i \le |C|,\ 1 \le j \le |C^*|.$$

It was chosen because it shows the harmonic mean of precision and recall and we can determine how good the performance of our approach is.

4.3 Definition of Baselines

In this section we define baselines constructed using (a) the widely used *k*-means algorithm [26], (b) a baseline based on the Topic Identification method for identifying the topics

included in a collection of weblog posts and a simple method for creating groups based on Jaccard coefficient similarity and (c) a baseline constructed by the standard Topic Identification method and grouped based on the most probable topics included in each document.

The first baseline was obtained by applying the standard *k*-means algorithm, with *k=10*, over the wide and narrow domain subsets. This state of the art algorithm is very well-known and can easily be compared with different approaches. The second baseline was constructed by applying the Topic Identification method from the original posts, i.e., without using the S-TEM methodology in the construction of the prototypes. In this second baseline the Topic Identification method was applied to the whole collection of weblog posts so we could get the topics most commonly discussed in the collection. We consider this baseline to be useful as it will provide a clear indication of the improvement that the S-TEM Methodology provides to our approach. Finally, the third baseline is basically a standard version of the Topic Identification model which estimates topics in each document and these documents are grouped accordingly to the top most probable topic of each document. This baseline was included only for comparison purposes as it is a standard reference of using the original Topic Identification method. The idea of the original approach is to identify the topics contained in each document. We have identified 10 topics (when possible) in a single document and grouped the documents based on the top most probable topic of each document. The topics (lists of keywords) were constructed with 10 terms each.

In Table 2, we present the F-Measure values of the three baselines presented. The best baseline result was generated with the standard *k*-means algorithm with 0.29 for wide domain and 0.24 for narrow domain. As we can see in Table 2 the improvement achieved with the wide domain dataset is much larger than the one obtained for narrow domain due to the fact

that the latter is limited to technology topics so that the topics and vocabulary are very similar.

We achieve broadly similar results with the second baseline, although we note a disimprovement in the case of the wide domain dataset. The third baseline reflects the impact of identifying the topics on individual weblog post and grouping them based on the most discussed topic in the weblog post. In some cases the most discussed topic in a post was not relevant for the whole collection and groups with few elements were created. In conclusion the poor information provided by the short text was an important factor.

4.4 Experiment 1: Unguided Initialisation in the Prototype Construction

In our first experiment, we do not provide a starting input in the process of Topic Identification. We have used the Topic Identification process with random initialisation, i.e., the LDA algorithm randomly selects posts (documents) that will be used as starting values for the probabilistic model; these posts are used for an initial estimation of the model. The algorithm will use them as the starting point to estimate the model as finite mixtures over an underlying set of latent topics (specialised distributions over words) inferred from correlations between words.

We obtained the best results when we selected 10% of the vocabulary to construct the prototypes, achieving average F-measure values of 0.44 and 0.28 for the wide and narrow domain datasets respectively. The rationale for limiting the vocabulary selected is to reduce the noise generated by the enriching technique (terms included in more than one category that can be highly correlated with discriminative information) and to highlight the most important features of each category. Table 3 presents a comparison of the approach presented against the baseline for the narrow and wide domain dataset. We have summarised the results in the

table showing the F-measure values obtained for a range of different prototype sizes (from 100 to 3000). We obtained best case values of 0.46 (wide domain) and 0.31 (narrow domain) with a prototype length of 2800 terms. We have also confirmed that in all the cases we have considerably outperformed the baseline.

We have limited the number of keywords selected in prototype construction from 100 to 3,000 terms per category in order to confirm the minimum number of terms needed for the prototype which can give us acceptable results in the clustering process. Although we carried out experiments with prototype length up to 3000 terms, we observed no discernible impact on the F-measure for prototypes length over 2100 terms. Furthermore, by reducing the number of terms, we can reduce the processing time, a fact which is particularly significant given the approach must be scalable to much larger corpora.

While significant improvements are recorded for both datasets, the gain achieved with the wide domain dataset is more significant than with the narrow domain. We consider that the reduced improvement in the latter case is due to the fact that when the enrichment process expands the corpus, it introduces some noisy terms, i.e., terms that share many categories in this kind of domain.

Although we had applied the Term Selection Technique to reduce this noisy information, it is difficult to highlight the discriminative information of each category. All of these considerations make the clustering task more difficult. Moreover, the size of the each document (in this case, weblog posts) is another important factor involved in this complex clustering process.

4.5 Experiment 2: Guiding the Initialisation of the Topic Model

In order to determine the impact on the results of the random choice of initialisation documents for the prototype construction, we have carried out experiments to establish the

degree of improvement which can be achieved when the Topic Model is provided with guidance. The key difference between these experiments and the previous one is that the LDA method is initialised with a "representative" document of each category. We present two variants of this experiment: firstly when the topics are selected manually by human experts, and secondly when a semi-automatic process is applied in the initialisation. We have proposed the initialisation for the Topic Identification process because we want to discard the possibility of allowing the algorithm to choose randomly the initial post, which may affect the quality of the final result.

In the first case, we have selected the initial prototypes (representative terms for each category) manually. A set of three human experts chose between 50 and 70 representative keywords for each category to be given to the LDA algorithm as initialisation documents. In the second case, the semi-automatic initialisation of the LDA model we have used the Transition Point technique which is based on the Zipf's law.

The general idea of this technique is that medium frequency terms are closely related to conceptual content in a collection. For this technique, we have used small set of weblogs documents randomly selected (3,000 posts in total) in order to identify the keywords of each category. We based our decision to select small numbers of weblog posts on the assumption that in many domains there are limited numbers of posts that can be used as a sample set and for this reason we proposed to seed the Transition Point technique with this limited number of weblog posts. Additionally, we do not consider it appropriate to use such a small dataset in a supervised approach due to the fact that learning algorithms do not perform well with very limited information. As it is well-known the larger the training set is the easier it is to find a good classifier [10] and typically the training runtime increases as the training set size increases. In addition we are dealing with weblog posts that tend to be short. Our purpose is

to provide an alternative for organising weblogs bearing in mind the data size restriction of information that could be found or provided, this aspect is predominant with weblog documents because of their constant dynamic changes on this kind of platform. The list of words obtained from either the Transition Point technique or the manual initialisation were used in the initial estimation of the topic models, in other words we provide starting points which may lead the inference procedure to the optimal parameters. Stop words and punctuation symbols were removed before the Transition Point technique was applied.

We present in Table 4 and Table 5 the results of the two approaches, applied to both the narrow and wide domain datasets. As before, we have varied the size of the prototype from 100 up to 3000 terms in order to establish the minimum number of terms needed for the clustering task.

The experiments using the wide domain dataset yielded better results, as the categories are better defined. The two approaches achieved very similar results although the best approach is when manual initialisation is applied to the Topic Identification process followed by very similar results from the semi-automatic approach. The best result (0.48 F-measure for both variants) was obtained with a prototype size of 1500 for manual guidance, representing both an improvement over the random initialisation, and is also significant in that it is achieved with a smaller prototype.

The results obtained when the approaches were applied to the narrow domain subset demonstrate improvement over the unguided approach, yielding highest F-measures of 0.35 and 0.34 for the manual and semi-automatic variants respectively. While lower than those achieved for the wide domain dataset, it still represents a significant improvement over the unguided approach.

5 Analysis of Results

As we have discussed previously the clustering of short, informal texts such as are found in weblog posts is a challenging task, however our approach has demonstrated that significant improvements can be achieved.

We have found that the best approaches are the prototype/topic based guided methods but we also achieved good results with the unguided version which is significant due to the fact that this approach do not require manual input or any external resource to initially guide the LDA model. We showed that the approaches presented in this work outperform the standard version of *k*-means, which we used as a baseline. We also demonstrated the valuable contribution made by the S-TEM methodology by defining a separate baseline in which this step was omitted.

In Table 6, we compare the results obtained in the two experiments and the three baselines already defined. In addition we have created an additional "reference" indicator in order to compare the results directly and to be able to see clearly the improvement that the expansion methodology provides to the clustering of weblogs. In this additional baseline we have employed S-TEM methodology before the standard *k*-means algorithm in order to provide a clear view of the benefit that S-TEM methodology provide to the clustering process. Basically, we employ the expansion methodology (S-TEM) to improve the representation of the data, after which we use *tf-idf* [36] to construct the similarity matrix which will be used in the *k*-means algorithm to construct the clusters, in order to achieve better performance. Finally, the clustering process of the datasets by the standard *k*-means clustering algorithm is performed.

In Figure 6 and Figure 7, we present a graphical representation of the results obtained by comparing each of the variations of the Topic Identification method and the *k*-means

algorithm. In these graphs we have included the experiments which use the Prototype/Topic Based Clustering and the baselines already defined (*k*-means and our baseline based on a Topic Identification method).

The experiments carried out use prototypes of different sizes (from 100 to 3000 terms) which are the basis of the P/TB Clustering process. The *k*-means algorithm is shown as a constant because it is not using the prototypes for the clustering process but it is included in Figure 6 and Figure 7 for comparison purposes. A clear improvement in the clustering process is achieved when the S-TEM methodology is applied in both domains to benefit the topic detection method (Manually, Transition Point and Unguided). S-TEM has provided relevant relations among terms with are beneficial to the clustering process. Although the improvement achieved with the narrow domain dataset is not as significant as for the wide domain dataset, we note the increased difficulty of the task in the former case. Regardless, we have significantly improved the baseline in all cases.

In Figure 8, we present a comparison between the average results obtained from our original experiments with those obtained by manually and semi-automatically selecting the initialisation documents for both datasets. It is clear that when we employed the wide domain dataset, the Topic Identification process has been aided by the enriching methodology as the categories are better defined, but the use of guidance does not result in much significant difference in cluster quality. On the other hand, when dealing with the narrow domain subset, the improvement is lower because of the high overlapping vocabulary. It is intrinsically difficult for the clustering process to define the boundary between the different classes in this type of case.

5.1 Results of Other Approaches

We have compared indirectly the results in proportion to the improvement obtained with the research work of [48] which uses an external resource. However we cannot compare directly the two approaches because the documents covered in this approach use different type of texts. We can see however that the authors achieved an improvement of approximately 5% from their defined baseline. In our case, although we do not use an external resource, we have achieved good improvement (more than 20% for wide domain and more than 6% for narrow domain).

We also compare indirectly the results of our research work with the clustering approach presented in [42]. The data used in this work are microblog documents and we have focused our experiments on weblogs. We have compared the proportional improvement obtained in this work with our results. The authors are only dealing with wide domain categories and they achieved up to approximately 20% of improvement in F-measure for the best cases compared to a standard *k*-means algorithm results. We have shown similar improvement for wide domain which is the domain used by Tsur, et al. In addition, we have shown good improvement when we are dealing with categories with high overlapping vocabulary. Finally, the authors have implemented a strategy to perform a fast clustering process; we believe that our clustering approach is simpler as our similarity measure is easy to compute.

It is not practical to compare our work directly against other research proposals using the same genre of text and the same aim of categorising weblogs based on topics discussed therein. For this reason we have compared our approach with the well-known *k*-means clustering method, thereby allowing the improvement of our approach to be seen clearly. We also proposed a baseline that shows the benefit of using the approach proposed in this research work.

Table 7 provides information of other approaches compared with our proposal. The contribution of our approach is that we provide the alternative of no using external resource for the clustering or expansion process. If limited information is available it may be used to initially guide the Topic Identification method.

In addition we have tested our approach in narrow domains that usually is a drawback for some of the approaches. Finally, the clustering process of using Jaccard coefficient is simple that is one of the points that we wanted to cover in the creation of the clusters.

6 Conclusions and Further Work

We have presented a novel methodology in which we analyse and organise short text. This methodology clusters weblogs based on a generative probabilistic model using predefined and non-predefined initialisation in conjunction with an enriching methodology. The methodology was applied to two different kinds of corpora, one considered as “narrow” domain with very similar categories, and the other one considered as “wide” domain with low overlapping vocabulary or dissimilar categories.

We have confirmed that our approach works well with wide domain corpora obtaining 0.48 in the best average F-measure value with just 10% of the vocabulary to generate the best prototypes. It has also shown improved results (albeit with a smaller gain) with narrow domains. Due to the simplicity of the clustering method used, the approach we have presented has shown acceptable ranges of processing times.

Finally, we have confirmed that the approach of random selection of initialisation documents works sufficiently well and that there is no statistical benefit to employing a more sophisticated method of input document selection. The results obtained were compared with a standard clustering algorithm.

In future work, we plan to modify the proposed approach and weight the expanded posts used in the generation of the prototypes with the aim of giving better information to the clustering process, thus improving the representation of the post, in particular in narrow domain environments. We are also interested in working on the scalability of the approach in order to be able to manage datasets with a large number of documents and classes. For this purpose, we are intending to adapt the approach described in [22] to tackle the problem discussed in this work.

Acknowledgments

The work of the third author was carried out in the framework of the WIQ-EI IRSES project (Grant No. 269180) within the FP7 Marie Curie, the DIANA APPLICATIONS Finding Hidden Knowledge in Texts: Applications (TIN2012-38603-C02-01) project and the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

References

[1] N. Agrawal, M. Galan, H. Liu and S. Subramanya, Clustering blogs with collective wisdom. *In Proc. of the International Conference on Web Engineering*, IEEE Computer Society, USA, 336-339, 2008.

[2] C. C. Aggarwal and C. Zhai, A Surver of Text Clustering Algorithms. *In A. C. Charu, & Z. ChengXiang, Mining Text Data*, Springer US, 77-128, 2012.

[3] C. C. Aggarwal and C. Zhai, A Survey of Text Classification Algorithms. *In C. C. Aggarwal, C. Zhai, C. C. Aggarwal, & C. Zhai (Eds.), Mining Text Data* (pp. 163-222). Springer US, 2012.

[4] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron and Y. Yang, Topic Detection and Tracking Pilot Study: Final Report. *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[5] J. Allan, R. Papka and V. Lavrenko, On-line new event detection and tracking. *In Proc. SIGIR International Conference on Research and Development in Information Retrieval*, ACM, NY, USA, 37-45, 1998.

[6] E. Amigo, D. Spina, B. Beotas, and J. Gonzalo, Towards an Evaluation Framework for Topic Extraction Systems for Online Reputation Management. *In Proc. of the Workshop on Dynamic Networks and*

Knowledge Discovery (DyNak), ECML/PKDD, Pensa, Cordero, Rouveroil, Troyano and Rosso (Eds.), CEUR-WS.org, Barcelona, Spain, 655, 2010.

[7] S. Banerjee and T. Pedersen, An adapted Lesk algorithm for word sense disambiguation using WordNet. *In Proc. of the CICLing 2002 Conference*, Lecture Notes in Computer Science, 3878, 136-145, 2002.

[8] I. Bhattacharya and L. Getoor, A latent dirichlet model for unsupervised entity resolution. *In the SIAM International Conference on Data Mining*, 2006.

[9] D. M. Blei, A. Y. Ng and M. I. Jordan, Latent Dirichlet Allocation. *The Journal of Marchine Learning Research*, JMLR.org, 3, 993-1022, 2003.

[10] D. Brain and G. I. Webb, On The Effect of Data Set Size on Bias And Variance in Classification Learning. *In Proc. of the Fourth Australian Knowledge Acquisition Workshop (AKAW '99)*. Sydney, Australia, The University of New South Wales, pp. 117-128, 1999.

[11] D. Boyd, A Blogger's Blog: Exploring the Definition of a Medium. *Reconstruction* 6(4), 2006.

[12] J. F. Cai, W. S. Lee, and Y. W. Teh, NUS-ML: Improving word sense disambiguation using topic features. *In Proc. of the 4th International Workshop on Semantic Evaluations (SemEval)*, Association for Computational Linguistics, Morristown, NJ, USA, 249-252, 2007.

[13] L. Fei-Fei and P. Perona, A Bayesian hierarchical model for learning natural scene categories. *In Proc. of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR, 2, 2005.

[14] C. Fellbaum, *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[15] C. Flynn and J. Dunnion, Topic Detection in the News Domain. *In Proc. of the 2004 International Symposium on Information and Communication Technologies*, ACM, 103-108, 2004.

[16] G. Grefenstette, *Explorations in Automatic Thesaurus Discovery*. Kluwer Ac, 1994.

[17] T. L. Griffiths and M. Steyvers, Finding scientific topics. *In Proc. of the National Academy of Sciences of the United States of America*, 101 (1), 5228-5235, 2004.

[18] T. L. Griffiths and M. Steyvers, A probabilistic approach to semantic representation. *In Proc. of the 24th Annual Conference of the Congnitive Science Society*, 2002.

[19] Z. Harris, *Distributional structure*. Word, 10 (23), 146–162, 1954.

[20] T. Hofman, Probabilistic latent semantic indexing. *In Proc. of the Twenty-Second Annual International SIGIR Conference*, ACM, pp.50-57, NY, USA, 1999.

[21] A. Hotho, S. Staab and G. Stumme, Ontologies Improve Text Document Clustering. In Proc. of the Third IEEE International Conference on Data Mining. Washington, DC, USA, IEEE Computer Society, 2003.

[22] R. M. Karp and M. O. Rabin, Efficient Randomized Pattern-Matching Algorithms. *IBM Journal of Research and Development*, 31(2), 249-260, 1987.

[23] V. R. Kumar and K. Raghuveer, Legal Documents Clustering using Latent Dirichlet Allocation. *International Journal of Applied Information Systems (IJAIS)*, 2(6), pp. 34-37, 2012.

[24] M. Lesk, Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. *In Proc. of the 5th Annual International Conference on Systems Documentation*. Toronto, Ontario, Canada, ACM, pp. 24-26, 1986.

[25] B. Li, S. Xu and J. Zhang, Enhancing Clustering Blog Documents by Utilizing Author/Reader Comments, *ACM Southeast Regional Conference*, 94-99, 2007.

[26] J. B. MacQueen, *Some methods for classification and analysis of multivariate observations*, Berkeley, University of California Press, 281–297, 1967.

[27] D. C. Manning and H. Schutze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.

[28] J. Peng, D. Yang, J. Wang, M. Wu and J. Wang, A Clustering Algorithm for Short Documents Based On Concept Similarity. *In Proc. of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing - PACRIM'07*. IEEE, pp. 42-45, 2007.

[29] F. Perez-Tellez, D. Pinto, J. Cardiff and P. Rosso, Characterizing Weblog Corpora. *In Proc. of the 14th International Conference on Applications of Natural Language to Information Systems*, NLDB-2009. Lecture Notes in Computer Science, Springer-Verlag, 5723, 299-300, 2009.

[30] F. Perez-Tellez, D. Pinto, J. Cardiff and P. Rosso, Clustering Weblogs on the Basis of a Topic Detection Method. In Proc. of the 2nd Mexican conference on Pattern recognition: Advances in pattern recognition, Springer-Verlag, Puebla, Mexico, 342-351, 2010.

[31] D. Pinto, H. Jimenez-Salazar and P. Rosso, Clustering Abstracts of Scientific Texts Using the Transition Point Technique. *In Proc. of the 7th International Conference, CICLing 2006*, Springer Berlin Heidelberg, 536-546, 2006.

[32] D. Pinto, P. Rosso and H. Jiménez, A Self-Enriching Methodology for Clustering Narrow Domain Short Texts, The Computer Journal, doi: 10.1093/comjnl/bxq069, 2010.

[33] A. Purandare and T. Pedersen, Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces. *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*. Boston, Massachusetts, USA, Association for Computational Linguistics, pp. 41-48, 2004.

[34] Y. Qiu and H. P. Frei, Concept based query expansion. *In Proc. of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, 160-169, 1993.

[35] H. Schütze, *Automatic word sense discrimination*. Computational Linguistics Journal, 24(1), pp. 97-124, 1998.

[36] G. Salton, A. Wong and C. Yang, *A Vector Space Model for Automatic Indexing*. Magazine Communications of the ACM. 18(11), pp. 613-620, 1975.

[37] K. Shubankar, A. Singh and V. Pudi, A frequent keyword-set based algorithm for topic modeling and clustering of research papers. *Data Mining and Optimization (DMO)*, 2011 3rd Conference on, Putrajaya: IEEE, 96-102, 2011.

[38] Y. Sekiguchi, H. Kawashima, H. Okuda, and M. Oku, Topic Detection from Blog Documents Using Users' Interests. *In Proc. of the 7th International Conference on Mobile Data Management*, 2006.

[39] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen, Short text conceptualization using a probabilistic knowledgebase. In Proc. of the Twenty-Second international joint conference on Artificial Intelligence. Barcelona, Catalonia, Spain, AAAI Press, pp. 2330-2336, 2011.

[40] J. K. Spärck, *A statistical interpretation of term specificity and its application in retrieval*, Journal of Documentation, University Press, 28, 11-21, 1972.

[41] M. Steinbach, G. Karypis, and V. Kumar, A comparison of document clustering techniques. *In KDD Workshop on Text Mining*, 2000.

[42] O. Tsur, A. Littman, and A. Rappoport, Efficient Clustering of Short Messages into General Domains. Boston. *The 7th International AAAI Conference on Weblogs and Social Media*. AAAI, 2013.

[43] C. J. Van Rijsbergen, *Information Retireval*. London: Butterworths, 1979.

[44] X. Wang and J. Wang, *A Method of Hot Topic Detection in Blogs Using N-gram Model*. Journal of Software, 8 (1), 184-191, 2013.

[45] C. Wartena and R. Brussee, Topic Detection by Clustering Keywords. *In Proc. of the 19th International Conference on Database and Expert Systems Application*, IEEE Computer Society, USA, 54-58, 2008.

[46] Y. Wilks, D. Fass, C. Guo, J. E. McDonald, T. Plate and B. M. Slator, *Providing machine tractable dictionary tools*. Machine Translation Journal, 5(2), pp. 99-154, 1990.

[47] D. Xing, and M. Girolami, *Employing Latent Dirichlet Allocation for fraud detection in telecommunications*. Pattern Recognition Letters, 28 (13), 1727-1734, 2007.

[48] T. Xu, and D. W. Oard, Wikipedia-based topic clustering for microblogs. *In Proc. of the American Society for Information Science and Technology*, American Society for Information Science and Technology, 48, pp. 1-10, 2011.

[49] C. Zhai, A. Velivelli and B. Yu, A Cross-collection Mixture Model for Comparative Text Mining. *In Proc. of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, WA, USA, ACM, pp. 743-748, 2004.

[50] G. K. Zipf, *Human Behaviour and the Principle of least Effort*. Addison-Wesley Press, 1949.

Tables

Table 1 Discussion Topics of the Two Datasets (narrow and wide domain).

	Category name	**Posts**	**Category name**	**Posts**
Subset 1 (Narrow Domain)	Cell_Phones_Plans	1,543	Video_Online_Games	6,578
	Computer_Networking	1,337	Maintenance_Repairs	1,973
	Programming_Design	2,466	Security	1,583
	Laptops_Notebooks	2,153	Music_Music_Players	1,640
	Software	4,800	Other_-_Internet	1,523
Subset 2 (Wide Domain)	Singles_Dating	20,498	Celebrities	2,219
	Software	4,800	Marriage_Divorce	2,956
	Womens_Health	4,262	Languages	1,914
	Politics	2,527	Elections	3,628
	Dogs	3,205	Books_Authors	2,468

Table 2. Baselines Applied to the Wide and Narrow Datasets.

Baselines	Wide domain	Narrow domain
Baseline a) K-means algorithm	0.29	0.24
Baseline b) Most similar prototype **(**LDA + Jaccard coefficient)	0.24	0.24
Baseline c) Standard Topic Identification method (LDA) applied to individual documents.	0.19	0.15

Table 3. Results of the Unguided Initialisation of the Prototype Construction (F-measure).

Prototype Size (No of terms)	Wide domain	Narrow domain	Prototype Size (No of terms)	Wide domain	Narrow domain
100	0.33	0.23	**1600**	0.44	0.28
200	0.35	0.23	**1700**	0.44	0.28
300	0.36	0.24	**1800**	0.44	0.28
400	0.36	0.26	**1900**	0.44	0.28
500	0.37	0.26	**2000**	0.45	0.28
600	0.4	0.27	**2100**	0.45	0.30
700	0.4	0.27	**2200**	0.44	0.29
800	0.41	0.27	**2300**	0.45	0.31
900	0.41	0.27	**2400**	0.45	0.31
1000	0.42	0.28	**2500**	0.45	0.30
1100	0.42	0.28	**2600**	0.45	0.30
1200	0.44	0.28	**2700**	0.45	0.31
1300	0.43	0.28	**2800**	0.46	0.31
1400	0.44	0.28	**2900**	0.45	0.32
1500	0.43	0.29	**3000**	0.45	0.31

Table 4. Manually Guiding the Initialisation of the Topic Model (F-measure).

Prototype Size	Wide domain	Narrow domain	Prototype Size	Wide domain	Narrow domain
100	0.37	0.26	**1600**	0.48	0.33
200	0.39	0.26	**1700**	0.48	0.33
300	0.40	0.27	**1800**	0.48	0.32
400	0.42	0.27	**1900**	0.48	0.32
500	0.43	0.29	**2000**	0.48	0.33
600	0.44	0.29	**2100**	0.48	0.33
700	0.45	0.30	**2200**	0.48	0.33
800	0.45	0.30	**2300**	0.48	0.33
900	0.45	0.31	**2400**	0.48	0.34
1000	0.46	0.31	**2500**	0.48	0.34
1100	0.46	0.31	**2600**	0.48	0.34
1200	0.47	0.31	**2700**	0.48	0.34
1300	0.47	0.31	**2800**	0.48	0.34
1400	0.47	0.31	**2900**	0.48	0.34
1500	0.48	0.32	**3000**	0.48	0.35

Table 5. Guiding the Initialisation of the Topic Model with the Transition Point Technique (F-measure).

Prototype Size	Wide domain	Narrow domain	Prototype Size	Wide domain	Narrow domain
100	0.36	0.25	**1600**	0.46	0.30
200	0.38	0.25	**1700**	0.46	0.32
300	0.38	0.26	**1800**	0.47	0.31
400	0.41	0.26	**1900**	0.47	0.32
500	0.42	0.28	**2000**	0.47	0.32
600	0.42	0.28	**2100**	0.47	0.32
700	0.43	0.28	**2200**	0.47	0.31
800	0.44	0.29	**2300**	0.48	0.33
900	0.45	0.28	**2400**	0.48	0.34
1000	0.45	0.28	**2500**	0.48	0.32
1100	0.45	0.28	**2600**	0.48	0.33
1200	0.45	0.28	**2700**	0.48	0.32
1300	0.45	0.29	**2800**	0.47	0.34
1400	0.46	0.3	**2900**	0.47	0.34
1500	0.46	0.29	**3000**	0.48	0.33

Table 6. Comparison of the Different Approaches to Cluster Weblogs (average F-measure).

Approach	Wide Domain	Narrow Domain
Experiment 1 Unguided initialization	0.44	0.28
Experiment 2 a) Initialisation manually guided	0.46	0.31
Experiment 2 b) Initialisation guided with TP technique	0.45	0.30
Baseline a) *k*-means algorithm	0.29	0.24
Baseline b) Most similar prototype **(**LDA + Jaccard coefficient)	0.24	0.24
Baseline c) Standard Topic Identification method	0.19	0.15
Reference Indicator S-TEM + *k*-means algorithm	0.42	0.25

Table 7. Comparison of approaches.

Approach	Dataset Type	Domain	External resource used	Improvement reported
Xu & Oard	Microblogs	Wide	Yes	5%
Tsur et al.	Microblogs	Wide	No	20%
Our proposal	Weblogs	Narrow	No resource and limited resource	6%
Our proposal	Weblogs	Wide	No resource and limited resource	20%

Figure Captions

Figure 1. The Prototype/Topic Based Clustering Methodology.

Figure 2 Clustering Results Using the Wide Domain Subset (average).

Figure 3 Clustering Results Using the Narrow Domain Subset (average).

Figure 4 Comparison of the P/Topic Based Clustering with and without Initial Guidance over Narrow and Wide Domain (best values).

Figures

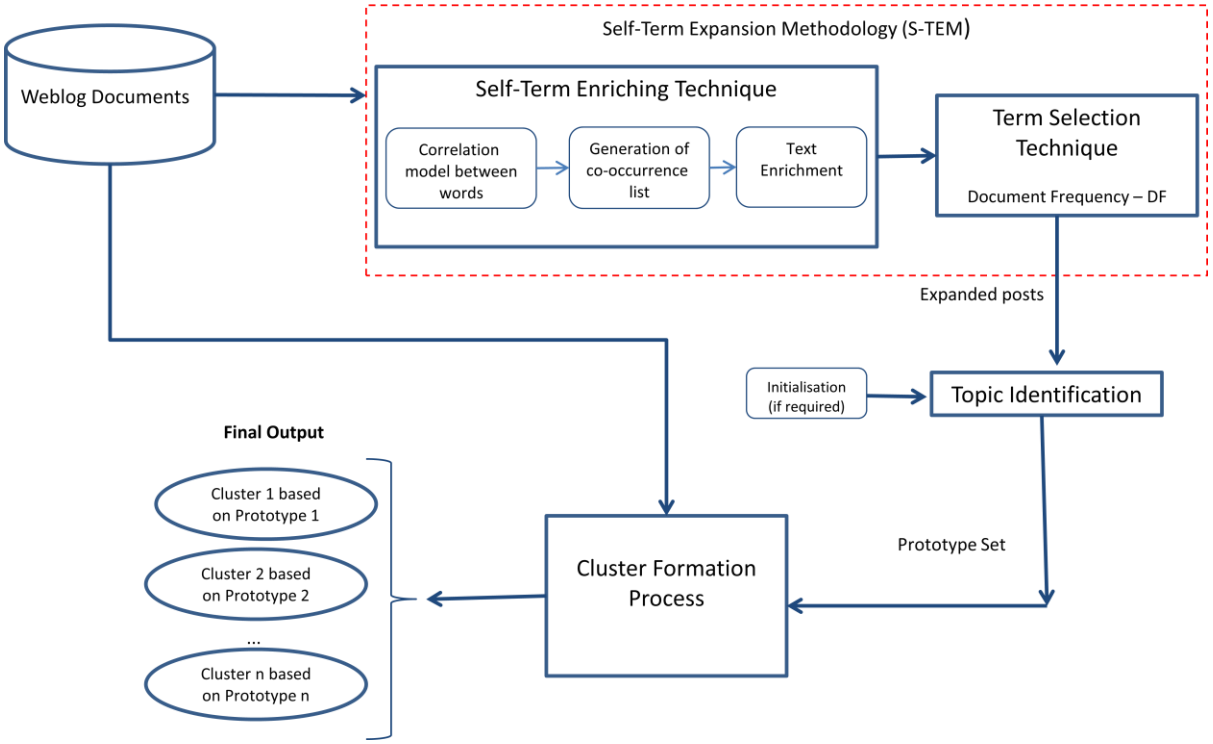

Figure 5. The Prototype/Topic Based Clustering Methodology.

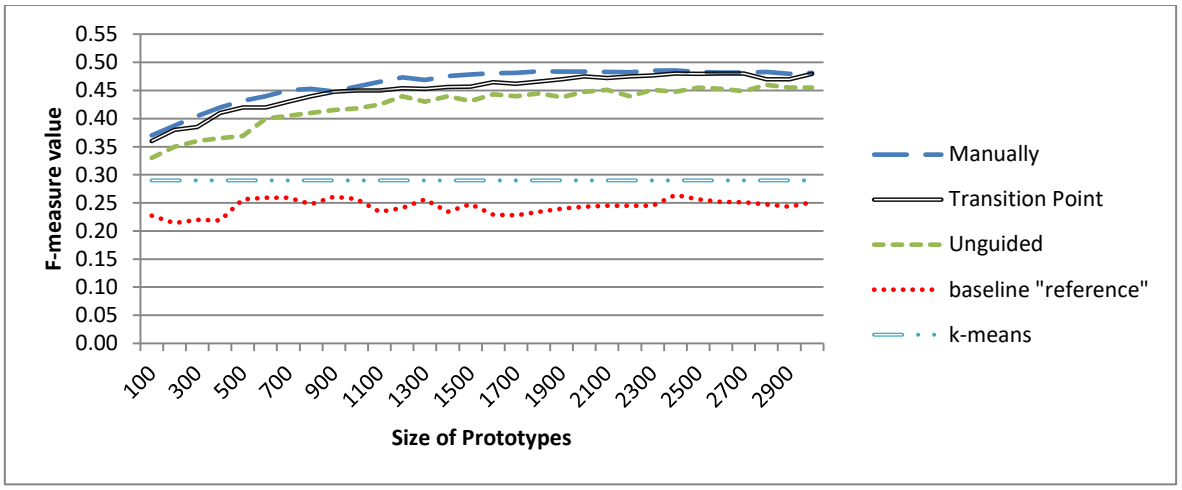

Figure 6 Clustering Results Using the Wide Domain Subset (average).

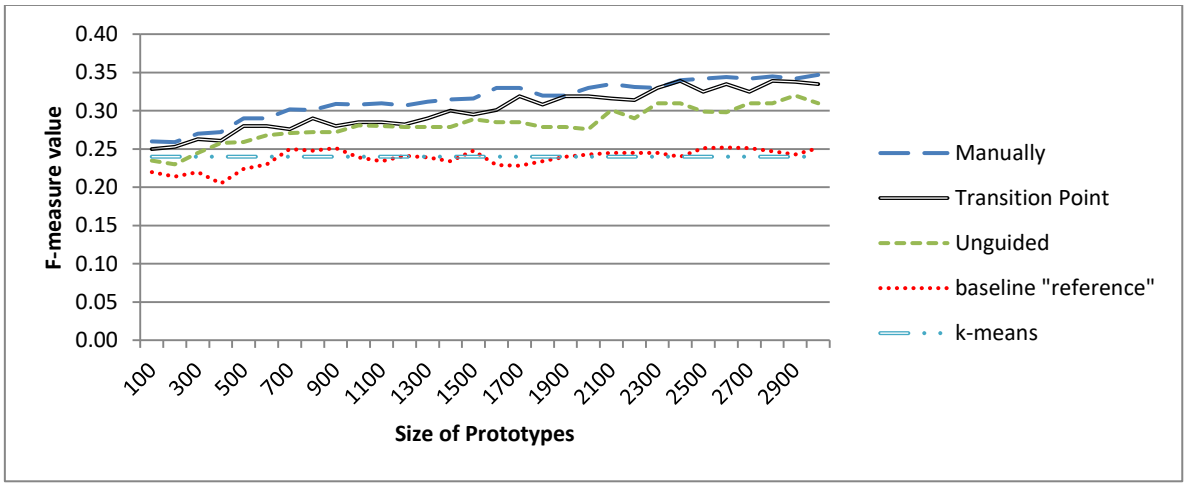

Figure 7 Clustering Results Using the Narrow Domain Subset (average).

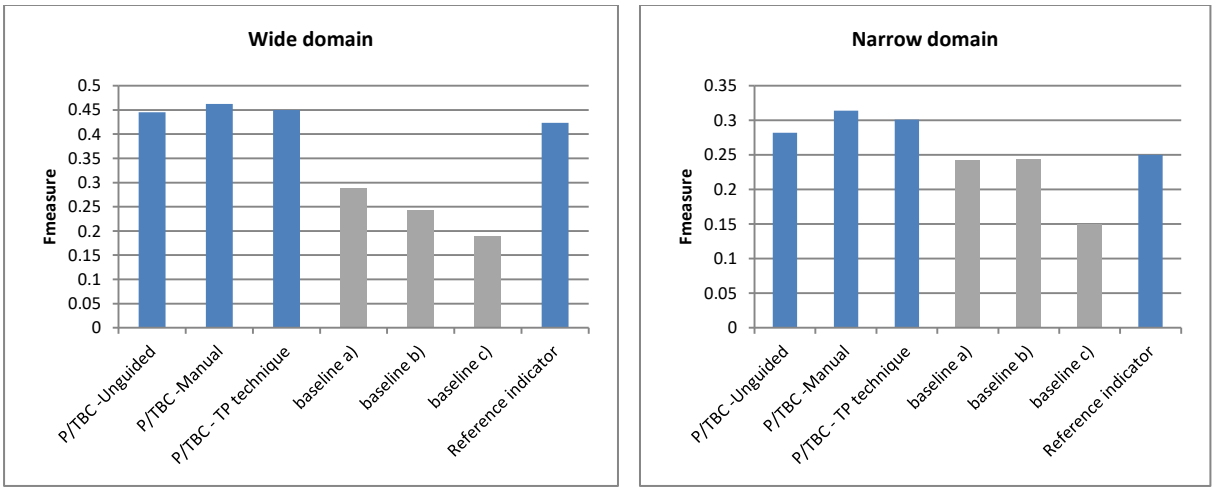

Figure 8 Comparison of the P/TB Clustering with and without Initial Guidance over Narrow and Wide Domain (best values).