

Automatic Topic Detection in Persian Blogs

Karine Megerdooian and Ali Hadjarian
The MITRE Corporation

ABSTRACT

User-generated communication such as blogs provide a powerful source for understanding cultural and religious values of a closed society, monitoring local attitudes, and helping to influence behavior. In this paper, we describe the development of a system that automatically detects the main topics and issues discussed in a community of Persian bloggers. The project combines language and text analytics with a study of the social network to build a model of blog clusters based on topic thus improving the capabilities of existing analysis tools on Persian-language blogs. In order to achieve the relevant text analytics for this project, we performed the first systematic study of the linguistic characteristics found in Persian language blogs and built a morphological analyzer able to process the word forms encountered in blog language. The project also developed a lexicon consisting of a growing number of loans and neologisms (or newly coined words) not included in existing systems. These tools were integrated within the system to build a topic classification tool that is able to identify the main topic of a blog post. This paper presents the development of these three components and an evaluation of the results.

DESCRIPTION

Understanding the human terrain is important in aiding intelligence analysis, developing strategic communications, and providing a meaningful context for on-the-ground military operations. User-generated communication such as blogs provide a powerful source for understanding cultural and religious values of a closed society, monitoring local attitudes, and helping to influence behavior. In this paper, we describe the development of a system that automatically detects the main topics and issues discussed in a community of Persian bloggers. The project combines language and text analytics with a study of the social network to build a model of blog clusters based on topic thus improving the capabilities of existing analysis tools on Persian-language blogs.

The language found on Persian blogs is often very conversational in style with substantial variance in orthography. Bloggers also use many borrowings from other languages or newly coined words that cannot be found in existing dictionaries. In order to achieve the relevant text analytics for this project, we performed the first systematic study of the linguistic characteristics found in Persian language blogs and built a morphological analyzer able to process the word forms encountered in blog language. The project also developed a lexicon consisting of a growing number of loans and neologisms (or newly coined words) not included in existing systems. These tools were integrated within the system to build a topic classification tool that is able to identify the main topic of a blog post. This paper presents the development of these three components and an evaluation of the results.

Several strategies were employed in extracting the entries included in the Blog Lexicon. Entries were automatically extracted and prioritized using information gain from a corpus of Persian language blogs or were collected manually from existing glossaries or through corpus

analysis. The current version of the lexicon contains 2,929 entries in 10 distinct categories and includes a collection of newly created terms and technical vocabulary with translations.

The collected blog posts are run through the Persian morphological parser that analyzes all word forms including compounds and provides a part of speech tag for the valid analyses [1]. The morphological formalism consists of a declarative description of rules utilizing typed feature structures with unification. The morphological analysis component takes advantage of a lexicon of about 40,000 entries in citation form. After morphological analysis, dictionary lookup eliminates all erroneous analyses. Any element that is not successfully associated with a word in the lexicon is tagged as an unknown. The current morphological analyzer has a coverage of 97% and an accuracy of 93% on a 7MB corpus collected from online news sources.

To automatically assign blog posts to a set of predefined topic categories, this study employs a profile-based classification technique based on an adaptation of the Rocchio algorithm, as proposed in [2]. Given the vector space representation of documents, the algorithm calculates a *prototype* vector, or a profile, for each document class. This prototype vector has the same dimension as the original document vectors. The weight of a given term in the profile is a combination of its weights in the positive and negative document vectors and is calculated using the following formula:

$$\omega_{c,k} = \max \left\{ 0, \beta \frac{\sum_{i \in \text{relevant}} \omega_{i,k}}{|\text{relevant}|} - \gamma \frac{\sum_{i \in \text{non-relevant}} \omega_{i,k}}{|\text{non-relevant}|} \right\}$$

where $\omega_{c,k}$ is the weight of the k^{th} term in the prototype vector for class c , $\omega_{i,k}$ is the weight of the k^{th} term in the vector representation of document i , *relevant* is the set of all positive training documents for class c , and *non-relevant* is the set of all negative training documents for class c . β and γ are parameters which control the contributions of relevant and non-relevant documents to the prototype vector, respectively. The resulting classifier, in essence, tries to determine the class of a given document based on its relative distance to the centroid of the positive and that of the negative examples [3].

Whereas in the original algorithm, the term frequency / inverse document frequency or TFIDF scheme is used to determine the weights of individual terms within each document (i.e., $\omega_{i,k}$ in the formula above), this study uses the entropy-based *information gain* (IG) measure as its weighting scheme of choice. The IG for each term is calculated by the following formula:

$$IG(c,k) = Entropy(c) - \frac{|c_k|}{|c|} Entropy(c_k) - \frac{|\overline{c}_k|}{|c|} Entropy(\overline{c}_k)$$

Where c is the set of all training documents for the given class, c_k is the set of all training documents for the given class in which term k occurs, and \overline{c}_k is the set of all training documents for the given class in which term k does not occur.

Information gain, unlike TFIDF, incorporates the class information to determine how well each term can discriminate between the positive and the negative document class. As such, it can be a more effective measure for calculating the term weights, when such class information is indeed available, as is the case here.

When trained on the terms (i.e., words) present in the documents for each blog topic category, the resulting profile-based classifier and the corresponding term weights can be used to identify the most significant key terms for each topic, with the most pertinent terms generally ending up with higher weights in such a classifier (Figure 1).

1	MEDICAL	11	MATERIAL
2	DOCTOR	12	PATIENT
3	PHYSICIAN	13	RESEARCH
4	TREATMENT	14	HYGIENE
5	REMEDY	15	BLOOD
6	CURE	16	AFFECTED
7	MEDICINE	17	MARROW
8	ILL	18	PULP
9	DRUG	19	DISORDER
10	ILLNESS	20	DERANGEMENT

Figure 1. The top 20 terms for the topic category Medicine

The category of a blog post can then be determined by the similarity of its vector representation to that of the prototype vector for each topic category. This similarity is often measured by the dot product of the two vectors. In situations where a crisp classification decision is desired, such a similarity can be compared to a pre-defined threshold to determine class membership. The value of such a threshold can be set to a value such as to maximize some performance criteria (e.g., classification accuracy) on the training data. The trained topic classification model was evaluated on a set of Persian documents from ten distinct topic categories. The documents were successfully classified with a range of 0.85-0.90 Area Under Curve measurements. The successfully identified topics were combined with a link structure analysis of a Persian blog community and we were able to automatically identify the main issues associated with each cluster.

BIOGRAPHY

Dr. Karine Megerdooomian is a theoretical and computational linguist specializing in less commonly taught languages, with emphasis on Persian and Armenian. She received her PhD in Linguistics from the University of Southern California in 2002 and has since worked as a computational linguist in the Computing Research Lab (CRL) in New Mexico, at Inxight Software in California, and at the Center for Advanced Study of Language (CASL) at the University of Maryland. She has also taught Persian and Armenian heritage language at UC San Diego. She is currently a Lead Artificial Intelligence Engineer in the Human Language Technology group at MITRE. Her current research focuses on Persian language blogs, especially with respect to analyzing BlogSpeak, identifying language change, and understanding public opinion.

REFERENCES

- [1] Amtrup, Jan W. 2003. Morphology in machine translation systems: Efficient integration of finite state transducers and feature structure descriptions. *Machine Translation*, 18(3), pp. 217-238.
- [2] Ittner, D.J., Lewis, D.D., and Ahn, D.D. (1995). Text categorization of low quality images. In Symposium on Document Analysis and Information Retrieval. Las Vegas, NV.
- [3] Sebastiani, F. 2002. Machine learning in automated text categorization. In *ACM Computing Surveys*, 34(1): 1-47