Contribution to research on new models of knowledge extraction for BigData systems

Héctor Cerezo-Costas, Advisor: F.Javier González-Castaño[1]

[1]Department of Telematics Engineering, University of Vigo

Motivation

Natural Language Processing (NLP) involves the application of mining algorithms to digital text content to extract knowledge. The interest in NLP has grown recently as there is more public media available in which users express needs, preferences and complaints. User reviews span multiple different domains and the text mining models must adapt to them quickly in real deployments. In a typical scenario, only a small subset of manually tagged data is available for a specific domain and large amounts of untagged data from the same and different contexts could complement it. Manual tagging is time-costly. Hence, there is an increasing need in innovative NLP algorithms capable of working with large amounts of data. Finally, NLP tasks may benefit from the storage and computing power of bigdata solutions.

There are many possible applications of this research such as product reputation analysis, online discussions mining and timeline thread analysis (thread discussions include emails and email-lists, internet forums, bulleting boards, etc), in order to characterize opinions on different topics and the roles of users in those discussions (user profiling).

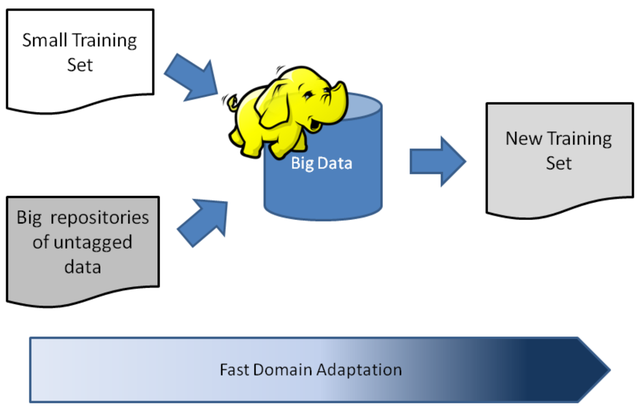

Figure 1 : Fast domain adaptation in NLP using public available data.

Thesis Objectives

- Research in new unsupervised algorithms for NLP tasks.
- Development of bigdata algorithms to solve NLP problems in terabyte-scale.
- Research in new technologies for fast adaptation of text mining models to different contexts.

Ongoing Work

Two main research lines:

1. Semi-supervised domain-specific connotation Graphs [1]
2. Real-time clustering at sentence-level [2]

Connotation Graphs

The aim of this algorithm is the discovery of polarity words that are domain-specific. In the process, a co-ocurrence graph is built and a ranking algorithm is applied. It is boosted by a small set of polarity words.

The algorithm works recursively. After each iteration new polarity words/combinations of words are encountered and they increase the seed in subsequent executions of the algorithm.

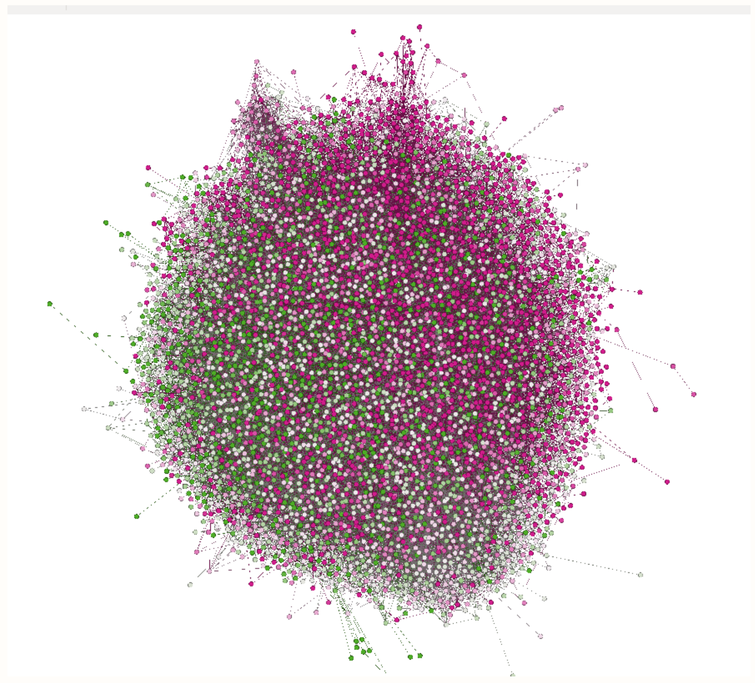

Figure 2 : Polarity graph after 50 iterations. Red and green nodes represent negative and positive words, respectively.

Techniques applied:

- Noise reduction by considering negation and sentences with opposite clauses.
- Insertion of bigrams and skipgrams in the graph, in order to obtaining domain-specific polarities [3]

Multiple applications to NLP tasks:

- Sentiment Analysis tools: increasing the reach of generic dictionaries
- User profiling: discriminating between recommendations, needs and complaints.

Real Time Clustering of Sentences

LSH is an unsupervised algorithm. In LSH the points are hashed. The probability of collision for nearby sentences in the R^N space is higher than for those that fall far apart. Many R^N spaces are considered in the process.

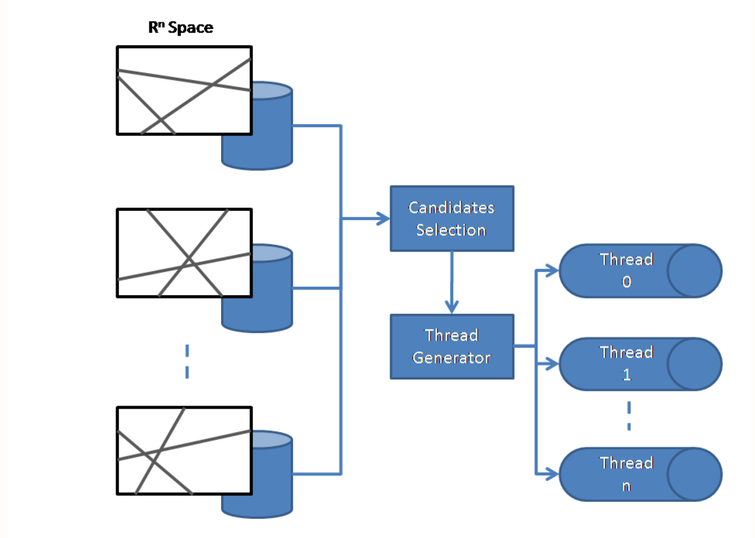

Figure 3 : Architecture of the system

Each sentence is vectorized and cosine similarity is employed to measure the closeness of two vectors.

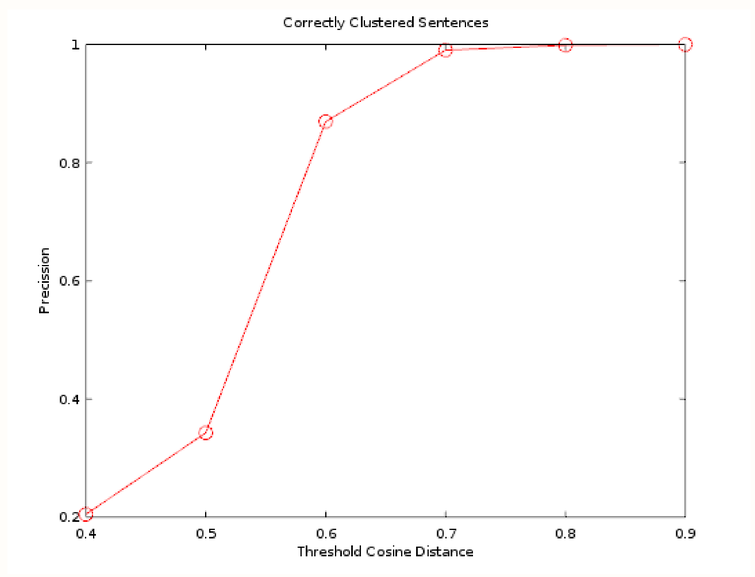

Figure 4 : Precision as a function of the cosine distance threshold (50000 records)

We are perforing tests at the moment with small sets to test the accuracy of the system. Dominant threads are manually tagged. The precision of the system grows as the cosine threshold increases.

The number of threads generated decreases with the cosine-distance threshold. There is a trade-off between precision and aggregation level.

Table 1 : Threads as a function of the cosine distance threshold (50000 records).

Cosine Threshold	Number of Threads
0.4	442
0.5	3868
0.6	18863
0.7	31131
0.8	34882
0.9	27967

Next Year Planning

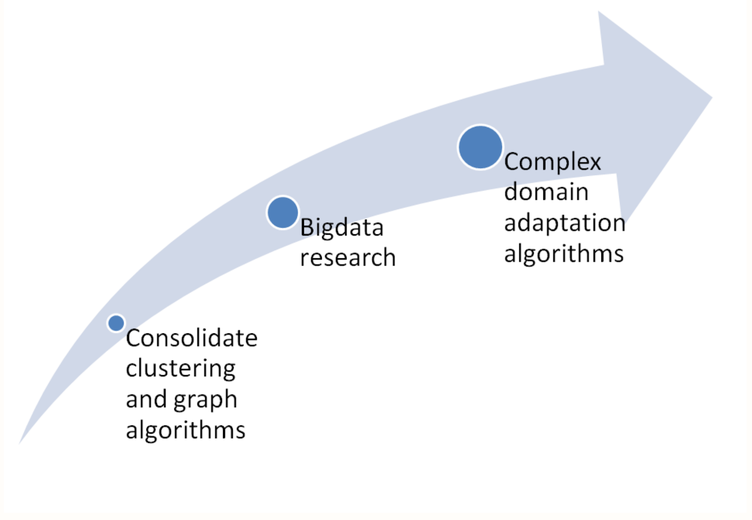

Figure 5 : Next year planning.

- Consolidation of current work. Generation of testing datasets to check more precisely the accuracy of the findings.
- Adapting the algorithms to bigdata platforms. Working with bigger graphs.
- Start working with domain adaptation of existing datasets with application to specific NLP problems (e.g. sentiment analysis [5], product recommendations, etc).

References

[1] S. Feng, R. Bose, Y. Choi. "Learning general connotation of words using graph-based algorithms", in *Proceedings of the Conference on Empirical Methods in Natural Language Processing 2011*, pp. 1092-1103. Association for Computational Linguistics

[2] N. Datar, N. Immorlica, P. Indyk, V.S. Mirrokni." Locality-sensitive hashing scheme based on p-stable distributions", in *In Proceedings of the twentieth annual symposium on Computational geometry 2004*, pp. 253-262. ACM.

[3] J. Fernández, Y. Gutiérrez, J.M. Gómez, P. Martinez-Barco, A. Montoyo, R. Munoz. "Sentiment Analysis of Spanish Tweets using a Ranking Algorithm and Skipgrams", in *Proc. of the TASS workshop at SEPLN 2013. IV Congreso Español de Informática*, pp. 17-20.

[4] A. Gionis, P. Indyk, R. Motwani. "Similarity search in high dimensions via hashing", in *Proc. VLDB 1999*, pp. 518-529.

[5] X. Glorot, A. Border, Y. Bengio. "Domain adaptation for large-scale sentiment classification: A deep learning approach", in *Proc. 28th International Conference on Machine Learning (ICML-11)*, pp. 513-520.