Comparing Talks, Realities and Concerns over the Climate Change: Comparing Texts with Numerical and Categorical Data

Yasir Mehmood, Timo Honkela
AaltoUniversity, School of Science
yasir.mehmood01@estudiant.upf.edu, timo.honkela@aalto.fi

Abstract

The conference on the climate change, UNFCCC 2010, took place from 29 November to 10 December 2010 in Cancun, Mexico. This paper presents an analysis of the opening speeches by various countries at the conference, combined with the statistics of the countries regarding their socioeconomic indicators and memberships of different climate treaties. A central objective is to compare different sources of the information that reflect the underlying complex system where there are obvious and less obvious relationships between the rhetoric and some aspects of reality. At the level of argumentation, we are interested in the occurrence of topics related to the climate change, i.e., whether some topics are mentioned or avoided in the speeches. The recognition of the topics is based on a semi-automatic term selection process that provides the input for the subsequent steps of the analysis. The data preparation process includes optical character recognition, machine translation and approximate string matching. We assume that the collection of terms serves as a relevant set of features that reflect the content of the speeches. These text-based features are then compared with the country statistics. The basic hypothesis is that there is a detectable but complex relationship between the content of the speeches and known facts. The most important contributions in this paper are the formulation of the basic questions and the overall hypothesis, an analysis of the relationships between the countries as well as between the topics and indicators, and the qualitative analysis of the results.

1. Introduction

The concerns over the climate change are growing particularly with the high emissions of greenhouse gases (GHG) that are raising the temperature of the planet. The United Nations Framework Convention on Climate Change (UNFCCC) is an international convention for the climate change. It came into force in 1994 with a mission of forming consensus among 30 developed countries to reduce GHG emissions into the atmosphere. However, since UNFCCC is a convention, its members do not have an obligation to guarantee a reduction in the GHG emissions. The legally binding framework of UNFCCC is Kyoto Protocol, which aims to reduce GHG emission by 5% from 1990s (see UNFCCC website http://unfccc.int/). The supreme body of Kyoto Protocol, responsible for the implementation of its aims, is called Meeting of the Parties to the Kyoto Protocol (CMP). Similarly the supreme body of the UNFCCC is called Conference of Parties (COP). In the opening session of both the COP and CMP, in UNFCCC 2010, all the member countries and organizations delivered their speeches to reflect their views on the climate change. In our analysis, the text of the speeches is the primary source of data. A secondary data consisting of statistics of the countries and their membership in climate treaties has been used to provide context to the primary data. This is further detailed later in this section.

1.1. Main Objectives

The analysis that we have performed is of two types. The aim of the first type of analysis is to visualize the data points on a two-dimensional grid. The data points are prepared from two data sources namely the primary and secondary data. The primary data is the speech text represented by a set of terms that include both unigrams and bigrams. The procedure for selecting the elements (or terms) is explained in section 3. The secondary data includes contextual information that is divided into two data sets namely *Country-wide Statistics* and *Membership of Climate treaties*. The *Country-wide Statistics* consist of quantitative information such as the GDP of a country, per capita income, mortality rate and various related facts. The *Membership of Climate treaties* includes information regarding a country's membership to treaties including Kyoto Protocol, Biodiversity, Desertification, etc. In our first analysis, the primary and secondary data are

combined by a scheme explained in Section 3.3, and the combined effect on the analysis is visualized by the Self-Organizing Map (SOM) [1]. The SOM is a widely used technique for analyzing and visualizing high-dimensional data. The SOM algorithm, positions each data vector, in the high-dimension, to a two-dimensional area without sacrificing the orientation of the vector in the original space. This results in a two-dimensional representation of complex data points where, the data points that are close to each other in the original space tend to remain close to each other in the two-dimensional space as well. This aids the visual perception of the data and hence one can identify the groups of similar data points. In the context of our analysis, the data points are the members of the climate change conference. Thus, visualizing a map of the (combined) data will unveil the similarity between/among different countries based on the contents of speeches, when contextualized with the respective country statistics. In addition to visualizing high-dimensional data, the SOM enables visualizing the distribution of each dimension - and therefore it can help in understanding the contribution of each data feature (textual or statistical) over the map. This is important, since the similarity in the speech documents is greatly influenced by the contribution of text [2] or statistics. Therefore, in the second part of our analysis, we select various features from the (combined) data and observe their distribution across the SOM map that was created in the first part of analysis. The SOM is one possible choice for conducting this kind of analysis but there are currently many other methods that could also be considered [3, 4]. However, the choice of methodology is not central aspect in our research, it is rather the formulation of the overall question and analysis architecture to increase understanding of the complex phenomenon.

1.2. Related Work

The topic of this paper is multifaceted and therefore providing a conclusive review of related work is challenging. Methodologically important aspect is the combination of text and data mining. There are a large number of scientific articles concerning text mining and data mining separately in domains like business and finance [5, 6] or biomedicine [7, 8]. The combination of text and data mining has been used much more infrequently. Existing work concentrates in the area of business and finance [9, 10]. The content of the negotiations have been widely reported in media and also analyzed in scientific articles in the areas of law [11] and environmental research [12].

This paper is organized so that Section 2 explains the data sources and its acquisition process. Section 3 details the data preparation process. Section 4 explains the results of experiments, and finally, Section 5 concludes the paper. In this article, the terms *speech*, *document* and *talk* refer to the same concept; similarly, the words *member* and *country* have been used interchangeably.

2. Data Acquisition

This section details the acquisition of primary and secondary data.

2.1. Primary Data

The primary data refers to the talks presented at UNFCCC 2010. There were 163 talks given by the heads of states and governments. We have managed to acquire 143 of them from the UNFCCC website since some of the talks are not available or easily translatable. The talks can be downloaded as PDF files and we have used Google's built-in OCR support for extracting the text from the PDF files. This process does not guarantee full accuracy of data acquisition since several *words* are distorted. On the other hand, we have avoided manual work to correct all the text documents considering it *a)* an interesting case of text mining, and *b)* time consuming human effort which can be addressed by smart methods. The details of dealing with this problem are explained in section 3. There are some talks that were presented in the national languages. We have used Google's translation facility in order to convert those in English. However, these have not been considered while creating the feature set. The textual data, acquired as described previously, is the primary data source for our experimentation because it has been given more weight in the analysis. However, the secondary source of data is explained in the next subsection.

2.2. Secondary Data

The secondary data, in our experimentation, refers to the contextual data. It is of two types: *Country-wide Statistics* and *Membership of Climate Treaties*, described earlier in Section 1. The contextual data has been acquired from CIA Factbook (https://www.cia.gov/library/publications/the-world-factbook/). The data preparation and representation is explained in Section 3.1. The contextual data has been given less weight in the analysis in order to amplify the impact of speech on the results of analysis.

3. Data Preparation and Representation

This section details the steps of data preparation as well as its appropriate representation for the analysis. Section 3.1 explains the preparation of primary data, Section 3.2 explains the preparation of secondary data and Section 3.3 outlines the final representation of data, which is necessary for the analysis.

3.1. Term Extraction, Validation and Weighting

In the primary data, the English text of the speeches is available for 106 countries out of 143. This is sufficient to create a collection of word features for the entire data set, without including the non-English speeches. The primary reason for excluding non-English text is that the Google translation facility requires human effort to understand the translation better [13] and therefore does not guarantee the exact translation. Moreover, if the individual terms, in the word features, are not chosen carefully, it may increase the feature space, giving rise to *sparsity* in the document vectors [14]. This has a great potential of affecting the quality of results by adding numerous features that are not meaningful to the analysis, and thereby adding to the computational complexity [15]. Feature extraction techniques such as principal component analysis (PCA) [16], Random Projection [17] and many others [18], can help overcoming this problem; however, these techniques represent the original feature space into a new space. This is not suitable for our analysis because we want to preserve the original (and meaningful) features in order to visualize the distribution of those features over the analysis map (as highlighted in the Section 1). On the other hand, a suitable feature selection technique can help alleviating this problem. Liu et al. [14] outlines various methods for reducing the original high-dimensional space by selecting useful features. In our work, we have employed *Entropy* to select the most informational features that comprise the set of terms (see also [19]). These features are essentially unigrams and bigrams having high entropy values. Considering each term vector a random variable T, the entropy H is calculated as follows:

$$H(T) = \sum_{t \in T} p(t) \log_2 p(t) \qquad (1)$$

The informational terms (unigram and bigrams) are selected by setting a threshold on the *H(T)*. Both the unigrams and bigrams obtained by this procedure are a little more than 1100 each. We have selected 400 unigrams with a little manual work. However, for the bigrams we have marginalized a lot of them based on their occurrences in a reference dataset (Europarl: http://www.statmt.org/europarl). This gives us nearly 200 bigrams and we select 35 most informational bigrams by a little manual selection.

Selecting the *set of terms* require an initial term-frequency matrix, comprising all the features (or term vectors). However, after selecting the features, based on entropy values, we perform a second pass on the whole document space in order to create a low-dimensional term-frequency matrix. In this pass, the term frequency is not calculated purely based on strict comparisons but with a slight tolerance for errors. The rational behind this is that the text extracted by Google OCR, misspells various words and frequent misspellings include missing letters or presence of some special characters inside the word. We have used a tool for approximate string matching called *agrep* (http://www.tgries.de/agrep/) for this purpose. The tolerance level was set to one edit error in the unigrams and two in the bigrams. The precision drops slightly by this method but the coverage of words or recall [2] is high. Finally, the term' frequencies are weighted by multiplying them with the inverse of documents in which they appear. This is commonly known as tf-idf representation. Finally, the term-frequency matrix obtained by the aforementioned procedure contains each document as a 435 dimensional vector D

$$D = [uni_1, uni_2, ..., uni_{400}, bi_1, bi_2, ..., bi_{35}] \qquad (2)$$

where, the first 400 dimensions represent unigrams and next 35 represent bigrams.

3.2. Information Extraction and secondary data

The preparation of secondary data mainly requires parsing the HTML pages of CIAFactbook (https://www.cia.gov/library/publications/the-world-factbook/) for both the *Country-wide Statistics* and *Membership of Climate Treaties*. This data has only been gathered for the countries that presented their speeches in UNFCCC. The data preprocessing includes mean-centering as well as scaling the values between 0 and 1. Table 1 shows a snippet of *Country-wide Statistics* that include 29 variables in total. The dataset for the *Membership of Climate Treaties* includes 26 different treaties (variables) for all 143 countries. Table 2 shows some of them. The numerical and categorical

Table I
A SELECTION OF COUNTRY-WIDE STATISTICS.

Countries	GDP Growth Rate	GDP Per Capita	Birth Rate	...	Pop. Below Poverty Line (%)
Guatemala	2.20	5200	26.96	...	56.20
Kenya	4	1600	33.54	...	50.00
Norway	1.50	59100	10.84	...	NaN
Germany	3.30	35900	8.30	...	11.00
Singapore	14.70	57200	8.50	...	NaN
Mexico	5.0	13800	19.13	...	18.20
...	...	...	...	...	...

Table II
A SELECTION OF THE MEMBERSHIP OF CLIMATE TREATIES.

Countries	Air Polution	Biodiversity	Climate Change	...	Whaling
Guatemala	0	1	1	...	1
Kenya	0	1	1	...	1
Norway	1	1	1	...	1
Germany	1	1	1	...	1
Singapore	0	1	1	...	0
Mexico	0	1	1	...	1
...	...	...	...	...	...

data shown in tables 1 and 2 respectively represent a 29 + 26 dimensional vector for each country.

$$D = [stat_1, stat_2, ..., stat_{29}, cat_1, cat_2, ..., cat_{26}] \quad (3)$$

3.3. Combining Text, Numerical and Categorical data

In order to carry out the analysis, the data from primary and secondary data sources has been combined. Since the secondary data, consisting of numerical and categorical values, provides context to the textual data, it is weighted less. Thus, the textual data will have more influence on the results of the analysis and this enables us to visualize the groups of countries on the SOM map primarily based on the content of the speeches presented in UNFCCC. Nonetheless, the real information regarding the countries, in the form of statistics, is present in the data providing less influential context to the data. The strategy to combine data from both the sources, to form a unified data V is shown in the following equation.

$$V = [D_1, ..., D_{435}] + \varepsilon[N_1, ..., N_{29}, N_{30}, ..., N_{55}] \quad (4)$$

The dimensionality of V is 490 and the value of ε is set to a small number in order to reduce the impact of numerical and categorical data on the analysis. The + sign in equation 4 does not signify an addition operation but combining both the datasets (textual + numerical and categorical).

4. Experimentation and Results

The experimentation is primarily of two types. In the first part of experimentation a SOM map of the entire dataset V, as shown in the equation 4, is created in order to see the position of data points (precisely countries/organizations in the conference) over the map. This map is shown in Fig. 1. In this figure, the shades of gray denote distances in the original high-dimensional data space. The darker the color the higher the distance is in the original space. The most interesting part of the map in Fig. 1, from the analysis point of view, is the top left and top center region. In this area most of the European and rich countries are in the proximity of each other. Moreover, various developing countries are scattered in small groups in the bottom of the map and a considerably visible group of some African countries can be found in the middle of the map.

The next part of experimentation deals with analyzing the distribution of various variables or components on the map. The first set of variables includes unigrams. These are 400 in total; the

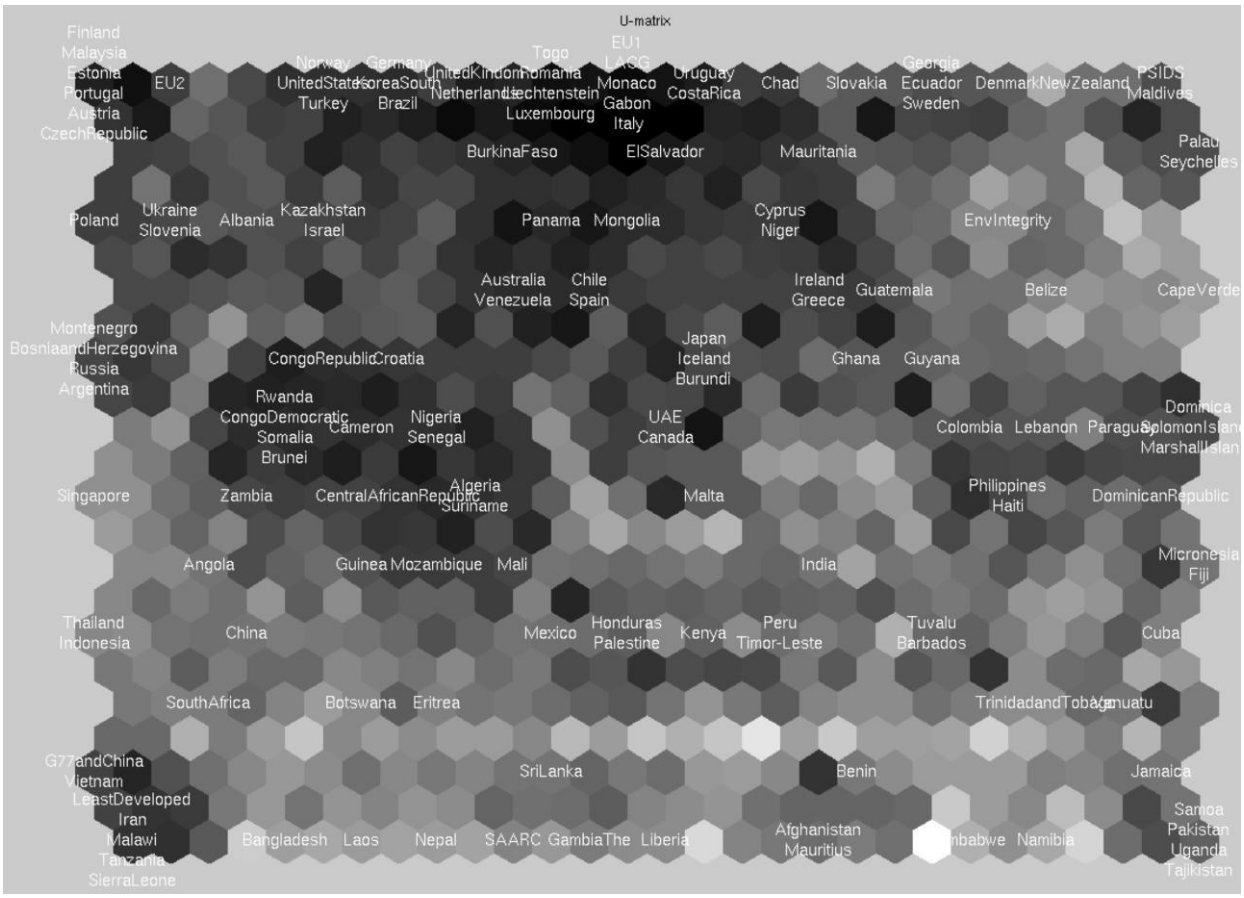

Fig. 1: SOM map of countries

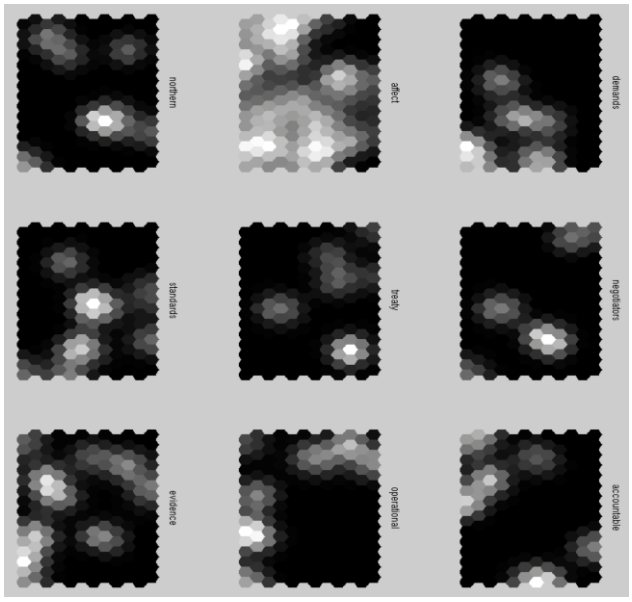

Fig. 2: Component map of unigrams

distribution of a few of them is shown in Fig. 2. The light shades of gray on the component maps show large values of the corresponding variable and dark shades show low values. Thus, we can see that in Fig 2 the unigram *affect* is unevenly scattered throughout the map except the top right. However, the top right portion is dark, explaining that countries in that region have used the word *affect* lesser than other countries. Moreover, we can also find interesting correlations between/among the unigrams. For instance, the density of *operational* is higher in those regions where the density is higher for *affect*. Interestingly this is the region where many underdeveloped countries are located as shown in Fig. 1. The next figure shows a component map of bigrams similar to the Fig. 2.

In Fig. 3 we can clearly see that the distribution of bigrams such as *climate change*, *Kyoto protocol*, *developing countries*, *legally binding* etc. is relatively higher in the regions of underdeveloped countries. This explains a correlation among these bigrams in the content of the speeches of underdeveloped countries.

The next two figures (Fig. 4 and Fig. 5) show the component maps of country-wide statistics and membership to the climate treaties respectively. Fig. 4 shows that *Life Expectancy at Birth* and *Literacy* is higher in the top regions of the map. This is clearly inline with the results shown in Fig. 1 since most of the developed world (including European countries and the US) is located in the top regions. Interestingly, these distributions do not deviate from the general understandings about the developed and developing world. In Fig. 5 we can see that significant portion of component maps of *Climate Change* and *Climate Change-Kyoto Protocol* are lighter in colors. This means that most of the countries have signed both of these treaties; however, some countries in the region of various African countries (see Fig. 1) have not signed *Climate Change*.

5. Conclusion

In this paper, we present a methodology to perform document analysis while putting contextual information in the background. The contextual information comprises of real statistics and their affect in the analysis is marginalized so that the analysis is dominated by the textual data. This has further helped us comparing the two sets of information and it reveals several interesting and obvious findings. We have found that the countries that are geographically close to each other and/or have similar socio-economic conditions are located in the proximity of each other on the analysis map. This is reinforced by analyzing the distribution of several variables, on the analysis map, representing concrete and real information regarding the countries. Our findings suggest that the countries that belong to similar groups in terms of their socio-economic conditions tend to speak in a similar manner when it comes to addressing the issue of climate change.

Our analyses open several areas of investigations for both social scientists as well as computer scientists. For example, it is worth asking if the divide between the rich, developing and under developed countries is also reflected in their actual concerns over the climate change. Or, as regards the global warming, do the underlying groups of countries (as investigated by this research) also take similar steps in order to address the challenges of climate change? Another simple step further in the context of this research is to take help from human experts to annotate the speeches (the data can be made available upon request) and then measuring the accuracy of the results presented in this paper. Finally, a concept level analysis of textual data, in a way that data features are essentially concepts (like ontologies in Semantic Web), would be an interesting investigation to consolidate the results of this papers in particular and document analysis in general. A by-product of such a research is a significant reduction in the size of lexicon and thereby the feature space [13, 20]. In general, we wish that the kinds of analyses presented in this paper could contribute in supporting sustainable developments in the world.

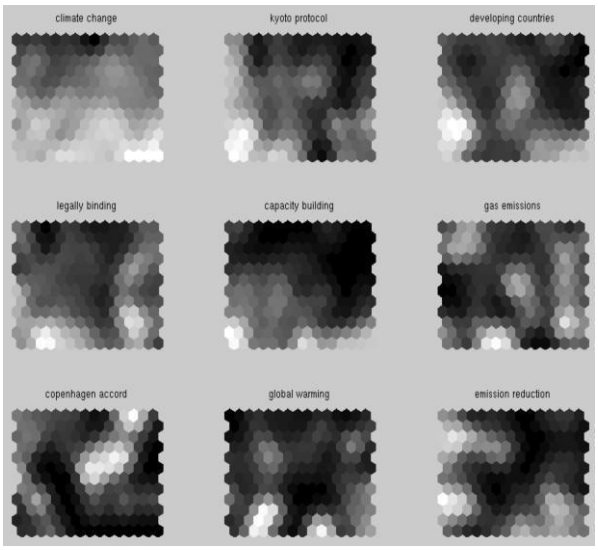

Fig. 3: Component map of bigrams

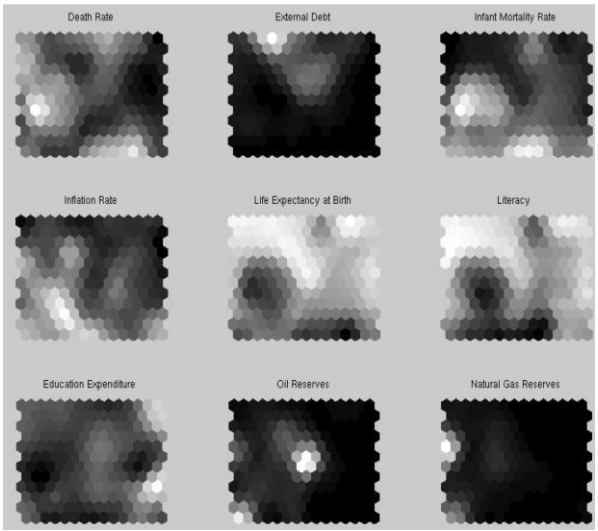

Fig. 4: Component map of numerical stats

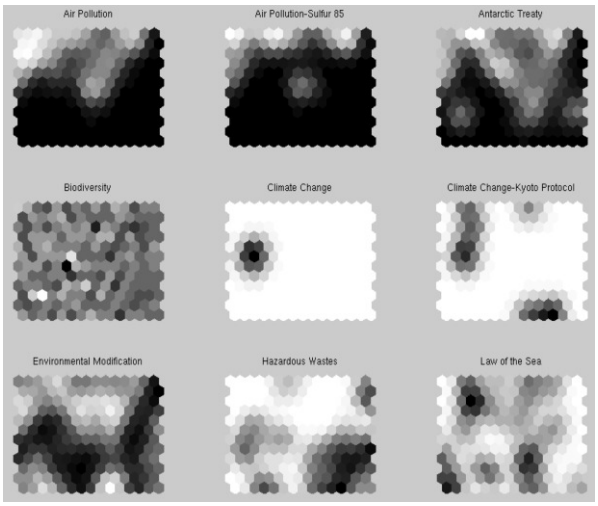

Fig. 5: component map of categorical data regarding climate treaties

References

[1] T. Kohonen. *Self-organizing maps.* Springer Series in Information Sciences, 2001.

[2] P. Senellart and V.D. Blondel. Automatic discovery of similarwords. *Survey of Text Mining II,* pages 25–44, 2008.

[3] M.A.A. Cox and T.F. Cox. Multidimensional scaling. *Hand- book of data visualization,* pages 315–347, 2008.

[4] P. Demartines and J. Hérault. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *Neural Networks, IEEE Transactions on,* 8(1):148– 154, 1997.

[5] G.J. Deboeck and T. Kohonen. *Visual explorations in finance with self-organizing maps,* volume 2. Springer, 1998.

[6] Debbie Zhang, Simeon J. Simoff, and John K. Debenham. Exchange rate modelling for e-negotiators using text mining techniques. In *E-Service Intelligence,* pages 191–211. 2007.

[7] Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B. Cohen. Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics,* 8(5):358–375, 2007.

[8] S Kaski, J Nikkilä, M Oja, J Venna, P Törönen, and E. Castrén. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics,* 4(1):48, 2003.

[9] A. Kloptchenko, T. Eklund, J. Karlsson, B. Back, H. Vanharanta, and A. Visa. Combining data and text mining techniques for analysing financial reports. *Intelligent Systems in Accounting, Finance & Management,* 12(1):29–41, 2004.

[10] S. Takahashi, M. Takahashi, H. Takahashi, and K. Tsuda. Analysis of stock price return using textual data and numerical data through text mining. In *Knowledge-Based Intelligent Information and Engineering Systems,* pages 310–316. Springer, 2006.

[11] D. Freestone. From copenhagen to cancun: Train wreck or paradigm shift? *Environmental Law Review,* 12(2):87–93, 2010.

[12] S.C. Walpole, S. Singh, and N. Watts. International climate negotiations: Health to the rescue? *The International Journal of Occupational and Environmental Medicine,* 2, 2011.

[13] P. Koehn. *Statistical machine translation,* volume 9. Cambridge University Press, 2010.

[14] T. Liu, S. Liu, Z. Chen, and W. Ma. An evaluation on feature selection for text clustering. In *MACHINE LEARNING- INTERNATIONAL WORKSHOP THEN CONFERENCE-,* volume 20, page 488, 2003.

[15] C.C. Aggarwal and P.S. Yu. Finding generalized projected clusters in high dimensional spaces. *ACM SIGMOD Record,* 29(2):70–81, 2000.

[16] I. Jolliffe. Principal component analysis. 2002.

[17] S. Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on,* volume 1, pages 413–418. IEEE, 1998.

[18] I.K. Fodor. A survey of dimension reduction techniques. *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory,* 2002.

[19] Mari-Sanna Paukkeri, Ilari T. Nieminen, Matti Pöllä, and Timo Honkela. A language-independent approach to keyphrase extraction and evaluation. In *Coling 2008: Companion volume: Posters,* pages 83–86, Manchester, UK, August 2008.

[20] A. Hotho, S. Staab, and G. Stumme. Ontologies improve text document clustering. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on Data Mining,* pages 541–544. IEEE, 2003.