INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

Detection and Geo-temporal Tracking of Important Topics in News Texts

Erik Michael Leal Wennberg

Dissertation for the achievement of the degree:

Master in Information Systems and Computer Engineering

Jury

Chairperson:	Prof. António Manuel Ferreira Rito da Silva
Supervisor:	Prof. Bruno Emanuel da Graca Martins
Co - supervisor:	Prof. Pável Pereira Calado
Member of the Committee:	Prof. David Manuel Martins de Matos

November 2011

Abstract

In our current society, newswire documents are in constant development and their growth has been increasing every time more rapidly. Due to the overwhelming diversity of concerns of each population, it would be interesting to discover within a certain topic of interest, where and when its important events took place.

This thesis attempts to develop a new approach to detect and track important events over time and space, by analyzing the topics of a collection of newswire documents. This approach combines the collection's associated topics (manual assigned topics or automatically generated topics using a probabilistic topic model) with the associated spatial and temporal metadata of each document, in order to be able to analyze the collection's topics over time with time series, as well as over space with geographic maps displaying the geographic distribution of each topic.

By examining each of the topic's spatial and temporal distributions, it was possible to correlate the topic's spatial and temporal trend with occurrence of important events. By conducting several experiments on a large collection of newswire documents, it was concluded that the proposed approach can effectively enable to detect and track important events over time and space.

Keywords: Newswire Documents , Important Events , Topic Modeling , Geo-temporal Topic Analysis , Latent Dirichlet Allocation

Resumo

Na nossa sociedade actual, a importância e o desenvolvimento de textos noticiosos tem tido um crescimento cada vez maior, o que reflecte as preocupações de cada população. Devido à grande diversidade de interesses de cada população, seria interessante descobrir dentro de um determinado topico de interesse, onde e quando estes eventos importantes ocorreram.

Esta dissertação de mestrado pretende desenvolver uma nova abordagem para detectar e rastrear eventos importantes ao longo do tempo e do espaço, através de uma analise temporal e espacial de uma colecção de textos noticiosos.

Esta abordagem combina os tópicos associados à colecção de documentos (tópicos anotados manualmente ou tópicos gerados automaticamente pelo um modelo probabilístico de tópicos) com os meta-dados espaciais e temporais associados a cada documento, para posteriormente analisar os tópicos da colecção ao longo do tempo utilizando series temporais, bem como sobre o espaço com mapas geográficos exibindo a distribuição geográfica de cada tópico.

Ao examinar as distribuições espaciais e temporais de cada tópico, verificou-se que a correlação entre as tendências espaciais e temporais dos tópicos e a ocorrência de eventos importantes. Foram realizadas algumas experiências a uma colecção de textos noticiosos, em que foi concluído que a abordagem proposta permite efectivamente detectar e rastrear eventos importantes ao longo do tempo e no espaço.

Palavras-chave: Textos Noticiosos, Eventos Importantes, Modelo de Tópicos, Análise Geotemporal de Tópicos, Latent Dirichlet Allocation.

Acknowledgements

The conclusion of this work represents a closing of another important chapter of my life. This was only possible thanks to the support of several people, of which I want to express my gratitude.

I would like to start to thank my parents, that always gave me all the opportunities and motivation to reach my life goals.

I also want to thank all my friends and colleges that were a fundamental key in my sucess and happiness, specially Vitor Loureiro, Marcos Simões, Maria Rosa, Marta Calvinho and João Soares.

Finally, I would like to thank my teachers, Professor Bruno Martins and Professor Pável Calado, without their constant guidance my work would never be possible.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the world of constant changes, news is in permanent development and an overwhelming amount of information is constantly distributed worldwide. Due to the wide diversity of interests of each population, it became an enormous challenge to discover where and when their important events took place, in order to discover the focus of interest of a given population in a moment of time.

The information available in the different newswire documents depends on the concerns shown by individuals or groups, which can be related to a particular time and space of interest. This relation can answer many interesting questions, such as: “Which are the main concerned topics referred in a geographic region?”, “How does a topic evolve over time and/or space?”and “What is the geographic distribution of a topic?”.

This thesis attempts to answer these questions, by performing a geo-temporal topic analysis on a large collection of newswire documents, in order to extract relevant spatial and temporal information related to events. However this thesis mainly focused on attempting to find a correlation between the spatial and temporal topic trends and the occurrence of important events, in order to detect and track these events over time and space.

1.1 Hypothesis and methodology

Given a large collection of newswire documents with a high topic diversification, it would be impracticable to manually search of where and when important events took place. Therefore, this thesis attempts to prove the following statements:

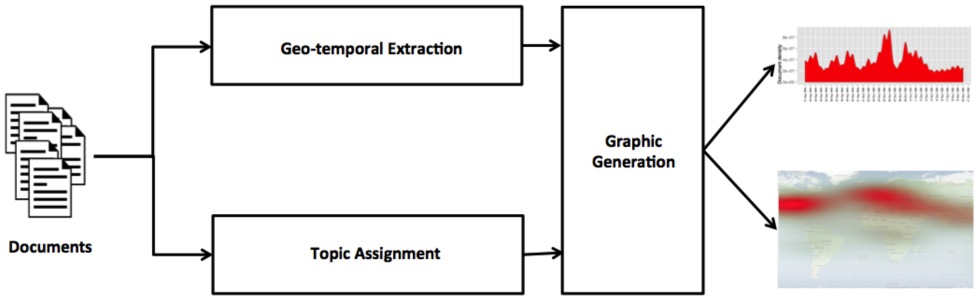

Figure 1.1: Proposed process pipeline of the approach's prototype.

1. **The geo-temporal topic analysis can effectively detect and track important events over time and space.**

2. **The generated topics from a Probabilistic Topic model are suitable for the geo-temporal detection and tracking of important events.**

In order to prove these statements, it was conducted a temporal and spatial analysis of the topics trends, which were associated with a collection of newswire documents (both manually assigned topics and automatically assigned and generated topics from a probabilistic topic model). This analysis sought to understand if indeed there is some kind of correlation between the topic's trends and the occurrence of important events. To this effect, it was developed a prototype that receives a collection of newswire documents, and in turn generates graphics displaying the temporal and spatial trends of each topic.

This prototype is based on the process pipeline of Figure 1.1, which is divided in 3 main tasks. Namely (1) geo-temporal extraction, which consists on extracting the spatial and temporal information of each document, (2) topic assignment, which classifies the collection of newswire documents in different topics, (3) and finally by combining the outputs of these two tasks, graphics are generated which are later used to analyze the collection's topics over time and space. Each of these tasks are described in more detail in Chapter 3.

1.2 Objectives and Contributions

The main objective of this thesis is to develop a novel approach to detect and track important events over time and space. Through the work, several experiments were conducted, of which derived several contributions, following being the most relevant:

- Proposal and evaluation of an approach to detect and track important events over time and space.
- Proposal of an evaluation method to validate the suitability of a topic to detect and track important events over time and space.
- Validation of the suitability of Latent Dirichlet Allocation generated topics to detect and track important events over time and space.
- Comparison of results of the geo-temporal topic analysis between using manual assigned topics and Latent Dirichlet Allocation generated topics.

1.3 Organization

The organization of the rest of this dissertation is as follows:

- Chapter 2 presents some fundamental concepts as well as reviews the most relevant related works.
- Chapter 3 describes in detail the different tasks involved in the proposed approach, as well as the software used and its respective configuration.
- Chapter 4 describes the evaluation methodology and discusses the obtained results of the proposed approach.
- Chapter 5 presents the main conclusions and contributions derived from this thesis. Additionally, this chapter ends by presenting some suggestions for future work.

Chapter 2

Concepts and Related Work

This chapter describes some fundamental concepts, as well as reviews the most relevant related work, specifically the Probabilistic Latent Semantic Analysis (pLSA) and the Latent Dirichlet allocation (LDA) models, as well as some extensions and applications in Topic Detection and Tracking.

2.1 Concepts

Topic Detection and Tracking (TDT) is an area of information retrieval which aims to (i) understand if a document created a new topic or if its topic was already referred in previous documents (i.e., First Story Detection), and (ii) recognize topics as described over documents (i.e., Topic Tracking) (Allan (2002)). Automatic approaches for TDT can be very useful, in order to analyze a collection of news reports. In brief, first story detection consists of recognizing, in an incoming stream of documents, which are the ones that describe a new topic. This task is often addressed by measuring the similarity between the respective incoming document and every document that appeared in the past. If the incoming document exceeds a similarity threshold with any one of the previous documents, then it is considered to represent an old topic otherwise it is considered to be about a new topic. It should be noted that the concept of new topic is relative, due the fact that a system can only base its decision on the previously analyzed documents. Topic Tracking consists of detecting a known topic in an incoming stream of documents. Each incoming document can be assigned a score referring to the matching between its contents and each of the considered topics. The best matching topic can also be obtained, and documents can be retrieved or filtered with basis on their topic matches. A document's score can be computed by measuring the similarity between the textual contents of respective document and a training set of documents for the topic.

In the context of TDT, as well as in most text mining and information retrieval tasks, documents are characterized by their words, also known as *terms*. This characterization can be made through different models, such as the *bag-of-words* model or topic distribution models. In the bag-of-words model, the only considered information is the frequency of each term, ignoring the order (Manning *et al.* (2008)). Typically, this model stores information regarding the frequency of each term in Document Vectors. Each component of these vectors corresponds to the occurrence frequency of a term and the dimensionality of the vector corresponds to the number of terms in the collection. If a term does not exist in a document, its value will be zero. Additionally, these values can be computed in different ways, depending on the weight given to different terms. A common approach is to use the term-frequency inverse document frequency (TF-IDF) heuristic (Manning *et al.* (2008)). The similarity of documents represented in the bag-of-words model can be computed through vector matching operations such as the cosine similarity (Manning *et al.* (2008)).

A Topic Distribution model represents documents as a mixtures of topics. The creation of each term is attributable to one of the document's topics. In natural language, words can be polysemous, meaning that words can have multiple senses. Consequently, the same word can belong to more then one topic. Therefore, the semantic ambiguity of words can only be resolved by other words given in the same context. Two of the most widely used topic distribution models are the probabilistic Latent Semantic Analysis (pLSA) model and the Latent Dirichlet Allocation (LDA) model (Steyvers & Griffiths (2007)). They will be presented in Section 2.2 of this document. The similarity of documents represented through topic distribution models can also be computed through the cosine similarity (i.e., documents are represented as vectors of topics) or through divergence functions such as the Kullback-Leibler divergence (Manning *et al.* (2008)).

In the context of TDT, a topic is generally viewed as a set of interconnected terms, where all terms are related to the same concept. For example, the terms *soccer*, *football*, *player* and *goal*, could be considered as belonging to the same topic. This view is aligned with that of probabilistic topic models, such as LDA and pLSA, which represent topics as mixtures of words (i.e., probabilistic distributions over terms). However, several previous works on TDT have reformulated the notion of Topic to *Event*, due to the fact that news reports often describe an unique happening, in some point of space and time. Therefore, one can conclude that an event can be associated to a combination of different contexted factors, such as a location or a temporal period referring to where and when it took place.

2.2 Related Work

2.2.1 Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis (pLSA), also known as probabilistic latent semantic indexing, is a technique that automatically indexes documents by topic assignments, based on a statistical latent topic model (Hofmann (1999)). The pLSA model extends the idea of Latent semantic analysis (LSA) (Deerwester *et al.* (1990)), by adding a probabilistic interpretation to the topics. Before describing pLSA, I will start by introducing the LSA technique.

In brief, LSA is an approach that maps documents and terms into a representation known as the Latent Semantic Space. This technique takes the vector space representation of documents based on term frequencies (i.e., a matrix of document vectors), and performs a low-rank approximation to this matrix, based on Singular Value Decomposition (SVD). The idea is that documents which show frequently co-occurring terms will have a similar representation in the latent space, even if they have no terms in common. Additionally, LSA performs a noise reduction as well as detects synonyms words that refer to the same topic.

Although LSA has achieved remarkable results in different domains, it still holds some problems, such as an incapability to deal with polysemous words or to explicitly distinguish between different meanings and different types of word usage. These limitations are due to the unsatisfactory statistical foundation. The pLSA model attempts to correct these issues, by creating a solid statistical foundation based on the likelihood principle, which will define a proper generative model for the documents.

The core of pLSA is a statistical model, called the *Aspect* model, which is a latent variable model for general co-occurrence data. The model assumes that a document d $\in D = \{d_1, ..., d_m\}$ in a collection of documents, and a word w $\in W = \{w_1, ..., w_n\}$ in a collection of words are conditionally independent given an unobserved topic z $\in Z = \{z_1, ..., z_k\}$ from a set of z topics. As illustrated in Figure 2.2, the model is represented in the plate notation, where the boxes or *plates* represent the sets. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. The white circles represent the unobserved variables and the gray circles represent the observed variables. This graphical model representation will be used for all the following models described in the report.

Translating the pLSA model to a probabilistic expression, leads us to the following:

$$P(d, w) = \sum_{z \in Z} P(z)P(w|z)P(d|z) = P(d) \sum_{z \in Z} P(w|z)P(z|d) \tag{2.1}$$

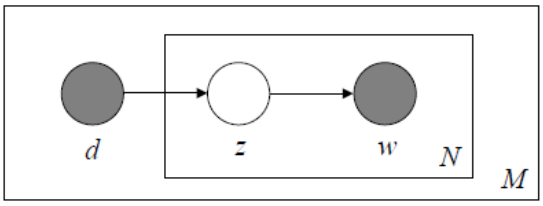

Figure 2.2: Graphical representation of the pLSA model using the plate notation (Hofmann (1999)).

In the equation, $P(z)$ is the probability of choosing topic z, $P(w|z)$ is the probability of a word w given the topic z and $P(d|z)$ is the probability of a document d given the topic z.

Following the likelihood principle, one can determine $P(d)$, $P(z|d)$, and $P(w|z)$, by maximization of the log-likelihood function

$$\mathcal{L} = \sum_{d \in D} \sum_{w \in W} n(d,w) log(P(d|w)). \tag{2.2}$$

In the formula, $n(d,w)$ represents the frequency of word w in document d.

In latent variable models such as pLSA, a standard procedure used for maximum likelihood estimation is the so-called Expectation Maximization (EM) algorithm (Dempster *et al.* (1977)). In brief, EM consists on two alternating steps:

1. An expectation (E) step, where the previous probabilities are computed for the topic z, based on the current estimates of the parameters:

$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{i=0} P(d|z_i)P(w|z_i)} \tag{2.3}$$

In the probabilistic expression, the probability of a word w in a document d, depends on the factor corresponding to a topic z ($P(z)P(d|z)P(w|z)$).

2. A maximization (M) step, where the model's parameters are updated according to the probabilities computed in the E-step

$$P(w|z) = \frac{\sum_d n(d,w)P(z|d,w)}{\sum_{i=0} n(d,w_i)P(z|d,w_i)} \tag{2.4}$$

$$P(d|z) = \frac{\sum_w n(d,w)P(z|d,w)}{\sum_{i=0} n(d,w_i)P(z|d_i,w)} \tag{2.5}$$

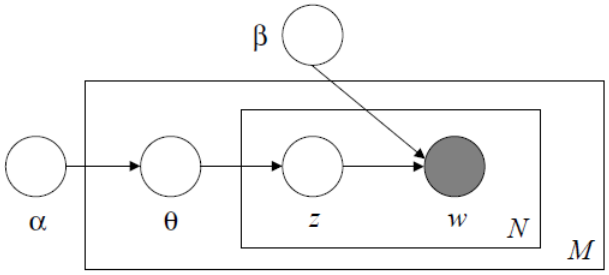

Figure 2.3: Graphical representation of LDA model using plate notation (Blei *et al.* (2003)).

$$P(z) = \frac{1}{\sum_{d,w} n(d,w)} \sum_{d,w} n(d,w) P(z|d,w) \tag{2.6}$$

These alternating steps define a convergent procedure that approaches a local maximum of the log-likelihood give in Equation (2.2).

2.2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for collections of discrete data, such as text documents (Blei *et al.* (2003)). The main idea of LDA is that documents can be represented as random mixtures of latent topics, where each topic is characterized by a distribution over words.

The LDA model, illustrated in Figure 2.3, assumes the following generative process for each document d in a corpus C:

1. For a document j,the model picks a value for the multinomial parameter θ_j of the vector $\theta_d = [\theta_{d1}...\theta_{dj}]^T$ over the N topics according to the Dirichlet distribution $\alpha = [\alpha_1...\alpha_n]^T$. The probability density function associated with a Dirichlet distribution returns the belief that the probabilities of K rival events are x_i given that each event has been observed $\alpha_i - 1$ times.

2. For a word i in document j, a topic label z_{ji} is sampled from the discrete multinomial distribution $z_{ji} \sim Multinomial(\theta_j)$.

3. The value w_{ji} of word i in document j is sampled from the discrete multinomial distribution of topic z_{ji} , which is generated from the Dirichlet distribution $[\beta_1...\beta_N]^T$ for each topic z_N.

For simplification reasons, let's assume (i) that the dimensionality k of the Dirichlet distribution is known and fixed, and (ii) that the word probabilities are parameterized by a k $\times$ V matrix β where $\beta_{ij} = p(w^j = 1|z^i = 1)$.

The LDA generative process can be sumarized in the following expression:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta), \tag{2.7}$$

In the formula, $p(\theta, z, w|\alpha, \beta)$ represents the probability of the joint distribution of a topic mixture θ, a set of K topics z, and a set of N words w given the parameters of α and β, where p($z_n|\theta$) is simply θ_i for the unique i such that $z_n^i = 1$. $p(\theta|\alpha)$ is the probability of θ given α and it is represented as followed:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1} \tag{2.8}$$

The probabilistic expression represents a probability density of the k-dimensional Dirichlet random variable θ. The parameter α is a k-vector with components $\alpha_i > 0$, and where Γ(x) is the Gamma function, which helps to interpolate the results.

If the expression is integrated over θ and summed over z, then we have the following:

$$p(d|\alpha, \beta) = \int p(\theta|\alpha)(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta))d_\theta. \tag{2.9}$$

In the equation, $P(d|\alpha, \beta)$ represents the marginal distribution of a document. Furthermore, taking the product of the marginal probabilities of single documents, we can obtain the probability of a corpus:

$$p(C|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d/\alpha)(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn}, \beta))d_{\theta d}. \tag{2.10}$$

Since LDA models the words in the documents under the *Bag-of-words* assumption (i.e, word order is not important and the occurrence of words is independent), it posits that the distribution of the words would be independent and identically distributed, conditioned on that latent parameter of a probability distribution. Thus, the words are generated by topics, and those topics are infinitely exchangeable within a document.

The probability of a set of words and topics takes the following form:

$$p(w,z) = \int p(\theta)(\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n))d_\theta, \tag{2.11}$$

In the equation, θ represents the multinomial parameter vector of the topics.

The main inferential problem that must be solved in order to use LDA is to compute a previous distribution of the hidden variables (θ_d, z_n, β) given a document w:

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}. \tag{2.12}$$

This distribution is intractable to be computed through exact inference procedures. Therefore, previous works have considered various approximate inference algorithms for LDA in order to compute these hidden variables, such as, variational EM (Dempster *et al.* (1977)) or Markov Chain Monte Carlo (MCMC) (Gilks & Spiegelhalter (1996)) methods such as Gibbs sampling (Blei *et al.* (2003)).

2.2.2.1 The Gibbs Sampling Procedure for LDA

The Gibbs sampling is a specific form of MCMC which simulates a high-dimensional distribution by sampling on lower-dimensional subsets of variables, where each subset is conditioned on the values of others. In the context of LDA, the procedure considers each word token in the document in turn, and the current word will be assigned with an estimated probability for each known topic. This estimation is conditioned by the topic assignments of all the other word tokens. From this conditional distribution, a topic is sampled and stored as the new topic assignment for this word. The distribution can be expressed as:

$$P(z_i = j|z_{-i}, w_i, d_i, .) = \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha} \tag{2.13}$$

where $z_i = j$ represents the topic assignment of token i to topic j, z_{-i} refers to the topic assignments of all other word tokens, and "." refers to all other known or observed information such as all other word and document indexes w_{-i} and d_{-i}, and the hyper-parameters α and β.

C^{WT} and C^{DT} are matrices of counts with dimensionality W×T and D×T respectively, where

C^{WT}_{wj} contains the number of times word w is assigned to topic j, not including the current instance i, and C^{DT}_{dj} contains the number of times topic j is assigned to some word token in document d, not including the current instance i. The Gibbs sampling procedure starts by assigning each word to a random topic in [1...T]. For each word, the count matrices C^{WT} and C^{DT} are first decremented by one for the entries that correspond to the current topic assignment. Then, a new topic is sampled from the distribution in Equation (2.13) and the count matrices are incremented with the new topic assignment. Each Gibbs sample consists on a set of topic assignments to all the N words in the corpus.

During the initial stage of the sampling process, the Gibbs samples have to be discarded because they are poor estimates of the posterior. After the *burn in* period, the successive Gibbs samples start to approximate the target distribution. At this point, to get a representative set of samples from this distribution, a number of Gibbs samples are saved at regularly spaced intervals, to prevent correlations between samples. This procedure can also estimate other hidden variables, such as $\theta^{(d)}_j$, which can be obtained from the count matrices as follows:

$$\theta^{(d)}_j = \frac{C^{DT}_{dj} + \alpha}{\sum_{k=1}^{T} C^{DT}_{dk} + T\alpha} \tag{2.14}$$

2.2.3 Temporal Latent Dirichlet Allocation

The Temporal Latent Dirichlet Allocation (TLDA) model is an extension of LDA which attempts to relate documents with the same topic that were written in different time periods (W.Reese & Shafto). This is a particularly hard challenge, due to the fact that a topic can change completly over time. For example, a medical document written in the 19th century might discuss bleeding or leeches, but current medical documents seldom mention these kind of treatments anymore. Even in the change of context within the field of the topic, one should nonetheless realize that both documents are related to the topic of medicine.

The TLDA model, illustrated in Figure 2.4, extends the LDA model by adding a K value, which adjusts a set of temporal splits to the document collection. This value can be infered by using a similar Gibbs Sampler to that which was previously described in Section 2.2.2.1. The parameter can be scored based on the probability:

$$P(k|z, w, \alpha, \beta) \propto \prod_{z,w} P(z|k, w, \alpha, \beta) \tag{2.15}$$

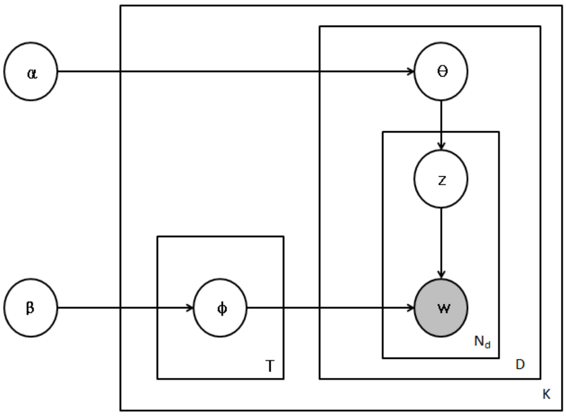

Figure 2.4: Graphical representation of the TLDA model using the plate notation (W.Reese & Shafto).

Using the Direchlet function described in Equation (2.8), it is desired to iterate separately over all of the topics and documents, and then calculate their respective Dirichlet score, as one can observe in the following equation:

$$P(k|z,w,\alpha,\beta) \propto \prod_{z,w}\left[\prod_{T}\frac{Dir(\langle w\rangle+\beta)}{Dir(\langle\beta\rangle)}\prod_{D}\frac{Dir(\langle\psi\rangle+\alpha)}{Dir(\langle\alpha\rangle)}\right] \tag{2.16}$$

In the formula, w represents the number of times a word as been assigned to a specific topic z, and ψ represents the number of words in a document assigned to a specific topic z. The previous equation allows one to sample the temporal split K in the documents-topics list.

2.2.4 The Geofolk Model

The Geofolk model is an extension of the LDA model for topic discovery using spatial information and word co-occurrences (Sizov (2010)). The model was originally applied to folksonomy resources such as tagged and georeferenced collections of Flickr[1] photos, although it can also be used over document collections.

The Geofolk model is given an arbitrary collection $D = \{d_1, ..., d_D\}$ of d documents, where each document d in this collection is composed of words 1...N_d (where $N_d \geq 1$). These words are taken from the vocabulary $V = \{w_1, ..., w_v\}$ that consists of V different words. Additionally, each document $d \in D$ is annotated with numeric attributes lat_d (latitude) $\in \mathbb{R}$ and lon_d (longitude) $\in \mathbb{R}$, which represent the spatial coordinates of the position of its creation.

[1] www.flickr.com

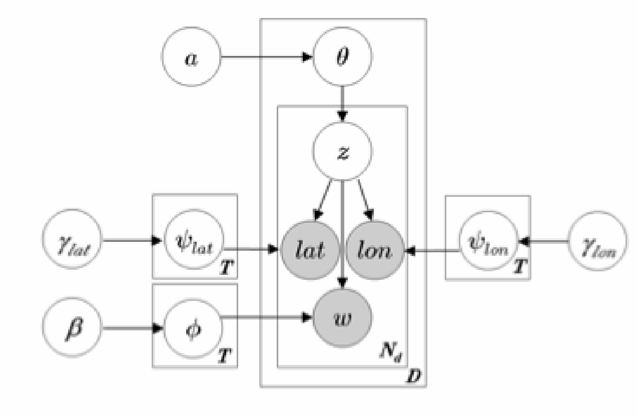

Figure 2.5: Graphical representation of the Geofolk model using plate notation (Sizov (2010)).

The generative process behind the Geofolk model for resources annotated with coordinates, illustrated in Figure 2.5, starts by executing LDA's steps as previously described in Section 2.2.2. In parallel, the topic generates two coordinates simultaneously, lat_{d_i} and lon_{d_i} from two topic-specific Gaussian distributions, ψ_z^{lat} and ψ_z^{lon} respectively. It is assumed that the Gaussian parameters μ_z^{lat} and μ_z^{lon} (i.e., the means for topic-specfic latitude and longitude) to compute ψ_z^{lat} and ψ_z^{lon} respectively, are drawn from a certain coordinate range using independent uniform distributions γ_{lat} and γ_{lon}. The normalization of spatial coordinates is necessary due, for example, two people take a picture of The Eiffel Tower, but each of them toke the picture from two different locations, such that the system must normalize the coordinates, so the photos can be associated to the same tag.

The Geofolk model can be applied to answer various questions about the similarity of resources and word co-occurrences, for which we will describe the most relevant applications for this thesis.

2.2.4.1 Keyword-based search with spatial awareness

In brief, this Geofolk application consists of treating a keyword-based query $q = w_{q1}...w_{qp}$ as a new annotated resource d_q with word co-occurrences $w_{q1}...w_{qp}$ and spatial preferences lat_q and lon_q, which can be transformed into the topical feature space of the Geofolk model. The required parameter estimation for θ_{d_q} is done by Gibbs sampling with previously learned and then fixed word co-occurrences distributions ϕ_z and spatial distributions ψ_z^{lat} and ψ_z^{lon} for all topics $z = 1...T$. For example, if a search for 'piccadilly' may be combined with coordinates of the London city center. This would help to filter out identically annotated but irrelevant resources such as pictures from the Manchester Piccadilly train station.

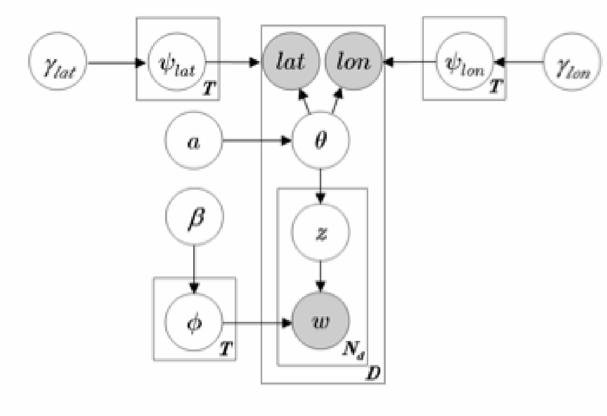

Figure 2.6: Graphical representation of the alternative Geofolk model for suggesting query locations using plate notation (Sizov (2010)).

2.2.4.2 Suggesting Locations for queries

Another interesting application of the Geofolk is the prediction of coordinates for keyword-based queries. For a keyword-based query $q = w_{q1}...w_{qp}$, its distribution over topics θ_{d_q} can be estimated together with the most likely coordinates through Gibbs sampling, when the Geofolk model is not conditioned on fixed values for lat_q and lon_q. Although this prediction of locations is ambiguous for queries that consist of more than one keyword. Therefore, it was developed an alternative generative process of Geofolk, which is better suited for this application. This alternative generative process consists on generating a single pair of values of lat_q and lon_q for a query q. The graphical model for this alternative process is shown in Figure 2.6.
The desired behavior can be achieved with Geofolk by importance sampling, from a mixture of per-topic Gaussian distributions, with mixture weights as the resource θ_d over topics. This distribution of coordinates remains parameterized by the set of coordinate-generating Gaussian distributions.

2.2.5 The Spatiotemporal Theme Model

The Spatiotemporal Theme Model is a topic modeling approach to analyze weblog topics and correlate the opinion of the weblog's author with when and where the weblog was written (Mei *et al.* (2006)). The collection of documents contains time stamps and location labels for each document, i.e., $C = \{(d_1, \tilde{t_1}, \tilde{l_1}), ..., (d_n, \tilde{t_n}, \tilde{l_n})\}$, where $\tilde{t_i} \in T = \{t_1, ..., t_{|T|}\}$ are the time stamps and $\tilde{l_i} \in L = \{l_1, ..., l_{|L|}\}$ are the location labels of document d_i. In addition, the approach introduces two concepts, namelly the Topic Life Cycle and Topic Snapshot.

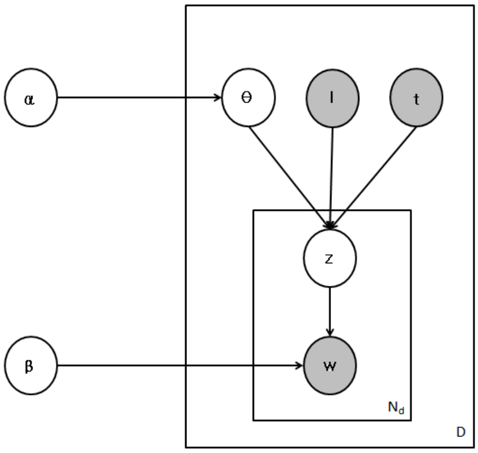

Figure 2.7: Graphical representation of spatiotemporal theme model using plate notation

1. Topic Life Cycle refers to a set of time stamps $T = \{t_1, ..., t_{|T|}\}$ where a topic z was referred in a location l. The Topic Life Cycle of a topic z at location l can be expressed as a conditional distribution $\{P(t|z,l)\}_{t \in T}$.

2. Topic Snapshot refers to a set of locations $L = \{l_1, ..., l_{|L|}\}$, a set of topics $Z = \{z_1, ..., z_n\}$, which were referred in a time stamp t. The Topic Snapshot at time t can be defined as a joint probability distribution of z and l conditioned on t, and can be expressed as $\{P(z,l|t)\}_{z \in Z, l \in L}$.

In a collection of documents C, the Spatiotemporal theme model (STM) considers the likelihood of a word w in a document d of time t and location l according to the mixture model given by:

$$P(w|d,t,l) = P(z_B)P(w|z_b) + (1 - P(z_B)) \sum_{j=1}^{k} P(w, z_j|d,t,l) \tag{2.17}$$

In the formula, $P(z_B)$ is the probability of choosing topic z_B.
STM estimates its parameters by using the Expectation-Maximization algorithm, which was introduced in Section 2.2.1.

Once all of the parameters are estimated, it is possible to obtain the Topic Life Cycle for a location $\tilde{l}$ by computing:

$$P(t|z_j, \tilde{l}) = \frac{P(z_j|t, \tilde{l})P(t, \tilde{l})}{\sum_{\tilde{t} \in T} P(z_j|\tilde{t}, \tilde{l})P(\tilde{t}, \tilde{l})} \tag{2.18}$$

In the formula $P(t, \tilde{l})$ is given by the word count in time period t at location $\tilde{l}$ divided by the total number of words in the collection. It is also possible to obtain the Topic Snapshot given the time stamp $\tilde{t}$ by computing:

$$P(z_j, l|\tilde{t}) = \frac{P(z_j|\tilde{t}, l)P(\tilde{t}, l)}{\sum_{\tilde{l} \in L} \sum_{j=1}^{k} P(z_j|\tilde{t}, \tilde{l})P(\tilde{t}, \tilde{l})} \tag{2.19}$$

With the Topic Life Cycles and Topic snapshots, various spatiotemporal patterns can be discovered and analyzed.

2.2.6 Topic Trend Discovery with LDA

Rzeszutek et. al proposed to combine the LDA topic model with a Self-Organizing Map (Kohonen *et al.* (2001)) to visualize topic trends over time in the Internet (Rzeszutek *et al.* (2010)). Since the Internet is a corpus which changes constantly over time, due to new documents being added to it, the set of descriptors produced by LDA from the N_D documents of the corpus can be seen as having a time component t associated to each document descriptor $\vec{\theta_d}(t) = [\vec{\theta_d} t]$, making it possible to analyze how the topics vary over time.

The arrival of new documents is very inconsistent, since the time between arrivals is more or less random, this complicates the analyses of the corpus, since most time-series methods assume a uniform sampling of the data. To resolve this issue, the document descriptors are analyzed by using a sliding window with a fixed step size. The size of the window T_{wnd} is defined as:

$$T_{wnd} = T_{end} - T_{start} \tag{2.20}$$

In the formula, T_{start} and T_{end} are the start and end times of the window respectively. Additionally, the article proposes to use a moving average approach for trend analysis. This approach consists on producing a document descriptor, which is the average descriptor for any particular time window, therefore representing the central tendency of the documents inside of that time window. For each window T_{wnd}, we obtain a descriptor $\bar{\theta}(T_{wnd})$ such that:

$$\bar{\theta}(T_{wnd}) = \frac{1}{N_D(T_{wnd})} \sum_{i=0}^{N_D(T_{wnd})-1} \vec{\theta_i}(T_{wnd}) \tag{2.21}$$

In the formula, $N_D(T_{wnd})$ is the number of documents inside of the time window at time T_{wnd}. The approach is very useful to analyze how topics vary over time, as it can be used to produce charts

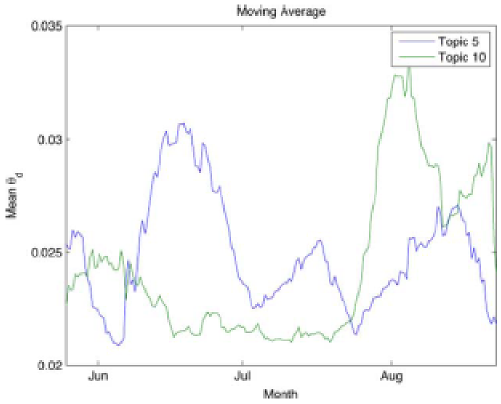

Figure 2.8: Moving average topic trend analysis for two topics (Rzeszutek *et al.* (2010)).

such as that of Figure 2.8. Nonethless, it becomes impossible to visualy analyze the topic trends if one attempts to analyze more then three topics. Thus the authors of the article proposed to use the Self-Organizing Map through a visual method to facilitate the visualization of the topic trends by mapping a high-dimensional feature space (i.e., the topic distributions for each document) onto a lower dimensional representation.

2.2.7 Temporal Text Mining

Text mining on online data could determine a number of important issues to many corporate functions, including brand monitoring, competition tracking, sentiment mining, and so on. Matthew Hurst proposed to observe the temporal pattern of a known term or class of documents using simple temporal models to determine which terms are trending in a given way in a time series (Hurst (2006)). A time series is a complete ordered sequence of periods, where each of them has a value. Given a time series T, a value of a time period i is represented by $t(i)$. Typically these values are normalized using a background time series T_{bg}. This time series represents the entire corpus of documents, all others being subsets of this corpus. Therefore the value of $t(i)$ after being normalized shall be equal to $t(i)/t_{bg}(i)$. To analyze a time series, one has to understand the different patterns involved in a time series. There are two simple elements that can be used to describe a time series, namelly $linear$ pattern which is a straight line and a $burst$ pattern which is described by being initially a flat line, followed by an acute jump in its last period. Each of these elements can be captured by a procedure which fits a model directly into a given time series. The procedures are the following:

Regression: used to return the linear regression components, such as the gradient m that allows us to classify the increasing and decreasing of a trend, as well as the r^2 correlation coefficient.

Burst: The score, $b(T, p, q)$, is computed as follows:

$$\frac{t(q)}{(\sum_{i=p}^{q} t(i))/(1 + q - p)} \tag{2.22}$$

In the equation, p and q represent respectively the first and last instant of the analyzed interval, and this equation captures the notion of a sudden jump in values, the ideal being a flat line with an infinite value in the final time period.

Each of these procedures allows to derive a number of metrics; some the most interesting are as follows:

Linear up : an upwards trend fitting a straight line, m must be greater than 0, the metric is $d(T) \times r^2$.

Linear down : a downwards trend fitting a straight line, m must be less than 0, the metric is $d(T) \times r^2$.

Final burst : a relatively stable trend with a burst in the final period, the metric is $b(T, 0, n - 1)$.

Maximum burst : $max(b(T', 0, p)) : p < n$.

In the metrics, $d(T)$ is the data density of series T and it is defined as:

$\sum_{0}^{n-1} 1$ if $t(i) > 0$; else 0/$|T_{bg}|$

The formula above represents the number of time periods with non zero values divided by the length of the background time series.

2.2.8 Interpolation Methods for Spatio-Temporal Data

Spatio-temporal interpolation methods (SIM) are required to estimate unknown values for unsampled location-time pairs. They are used, for instance, to provide contours for displaying data graphically (i.e., estimate some property of the surface at a given location) (Demirhan *et al.* (2003); Li & Revesz (2004)). These estimations are based on the assumption that spatially closer locations are more likely to have similar observation values than those which are far apart. Nonetheless, there are usually uncertainties associated to these estimations due to the fact the observed points give an indication about the surface's character at their corresponding locations, but not about the whole surface. Therefore, these methods usually assume that the value of a data point is valid for the area covered by the grid in which it lies. There are two different approaches for spatio-temporial interpolation, namelly (i) reduction, which reduces the spatio-temporal problem to a regular spatial interpolation and (ii) extension, which deals with time as another dimension in space and extends the spatio-temporal interpolation problem into a higher dimensional spatial interpolation problem.

There exist several different interpolation methods, such as Radial basis functions, kriging, the minimum curvature method or inverse distance weighting (Li & Revesz (2004)).

2.2.8.1 Inverse Distance Weighting

The inverse distance weighting (IDW) interpolation method is based on the assumption that things that are close to one another are more alike than those that are farther apart (Li & Revesz (2004)). Therefore the observed samples closer to the estimated unbsorved locations will have more influence than on those farther away. The IDW assumes that each measured point has a local influence that diminishes distance, thus, near neighboorhood are given high weights, whereas points at a far distance are given small weights. The general formula of IDW interpolation is the following:

$$w(x, y) = \sum_{i=1}^{N} \lambda_i w_i \tag{2.23}$$

In the formula, $w(x, y)$ is the predicted value at location (x, y), N is the number of nearest known points surrounding (x, y), λ_i are the weights assigned to each known point value w_i at location (x, y) and can be computed as follows:

$$\lambda_i = \frac{(\frac{1}{d_i})^p}{\sum_{k=1}^{N} (\frac{1}{d_k})_k^p} \tag{2.24}$$

In the formula, d_i are the Euclidean distances between each (x_i, y_i) and (x, y), and p is the exponent, which influences the weighting of w_i on w.

2.2.8.2 Kriging

Kriging is a popular interpolation method used by practitioners in fields such as geographical mapping (Demirhan *et al.* (2003)). The idea behind this method is to consider all observations as a realization of a random spatial process. There are several versions of kriging, such as simple Kriging, ordinary Kriging and universal Kriging, which result in different topologies. I will now describe the steps taken to generate the interpolated grid using the simple kriging. First, the method analysis and expresses the spatial variability with a variogram function, γ, which can be expressed according to the following semi-variance equation:

$$\gamma(h) = \sum_{i,j \in H} \frac{[f(z_i) - f(z_j)]^2}{2N(h)} \tag{2.25}$$

In the equation, $N(h)$ is the number of observations separated by a distance h, and H is the set of observations h distance apart. The above experimental variogram is used to construct a theoretical one by applying linear least squares method and thus estimating the parameters of the theoretical variogram and $f(z_i)$ is the linear unbiased predictor function. Given the theoretical variogram, now the method has to search for the best linear unbiased predictor to achieve the value of an unobserved location based on the observed locations.

Let us consider the unbiased estimate $f(z)$ from the neighboring data values $f(z_i)$, where $i = 1, ..., n$. The model $f(z)$ is stationary with mean m and covariance $C(h)$. The covariance is related to the semi-variogram according the following expression:

$$2\gamma(h) = Var[f(z_i + h) - f(z_i)] = 0.5[C(0) - C(h)] \tag{2.26}$$

where $C(0)$ is the stationary variance.
The simplest form of kriging, so called as simple kriging, considers the following linear estimator:

$$f'(z) = \sum_{i=1}^{n} \lambda_i(z) f(z_i) + [1 - \sum \lambda_i(z)]m \tag{2.27}$$

In the equation, f' is the linear regression estimator and m is the known mean value. The simple kriging weights $\lambda_i(z)$ are determined to minimize the error variance.
The minimized estimation variance or kriging variance is expressed as followed:

$$\delta^2(z) = C(0) - \sum_{i=1}^{n} \lambda_i(z)C(z - z_i) \geq 0 \tag{2.28}$$

2.2.9 Kernel Density Estimation

Kernel Density Estimation (KDE) is a non-parametric method of extrapolating point data over an area of interest without relying on fixed boundaries for aggregation (Carlos *et al.* (2010)). The density of points is calculated using specified bandwidth (a circle of a given radius centered at focal location). This produces a smooth, continuous surface where each location in the study area is assigned a density value, which can then be used as the independent or dependent variable in statistical models. Using the point density function, it is possible to calculate the density value of each point. This function defines the number of cases per unit area at each location throughout an area of interest. To calculate this density surface, for each case, a $neighborhood$ is delineated, usually by defining a search radius. The points that fall within this radius are divided by the area of the $neighborhood$. The point density function is defined as:

$$\lambda(x, y) = \frac{n}{|A|} \tag{2.29}$$

In the equation, $\lambda(x, y)$ is the point density at location (x, y), n is the number of events and $|A|$ is the area of the $neighborhood$. When $neighborhoods$ overlap, the results are summed to indicate a higher density of cases. The units of $\lambda(x, y)$ are cases per unit area. It is to be noted that the point density function does not consider the spatial configuration of features of interest within the bandwith. Therefore all the locations within the $neighborhood$ radius will have the same density value, which is unlikely to happen. In order to compensate, a density function can incorporate a decay function to assign smaller values to locations which are still in the $neighborhood$, but more distant from a case. This approach is contemplated by KDE. There are two different KDE approaches, namelly (1) the static bandwidth KDE, which fits a curved surface over each case such that the surface is highest above the case and zero at a specified distance (bandwidth) from the case and (2) the adaptive bandwidth KDE which uses a bandwidth based on a geographic distance.

The static bandwidth KDE's density value of each location is calculated as follows:

$$f(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} K(\frac{d_i}{h}) \tag{2.30}$$

In the equation, $f(x,y)$ is the density value at location (x,y), n is the number of cases, h is the bandwidth, d_i is the geographical distance between case i and location (x,y) and K is a density function which integrates to one. The units of $f(x,y)$ are cases per unit area.

The adaptive bandwidth KDE method uses background population drawn from LandScan data to calculate a kernel of varying size for each individual case (Dobson *et al.* (2000)). The landscan is an algorithm which uses spatial data and imagery analysis technologies and a multi-variable dasymetric modeling approach to disaggregate census counts within an administrative boundary. This limits the influence of a single case to a small spatial extent where the population density is high as the bandwidth is small. The density value of each location is calculated as follows:

$$f(x,y) = \sum_{i=1}^{n} K(\frac{d_i}{p(u,v)}) \tag{2.31}$$

The main diference between this equation and Equation 2.30 is the usage of diferent types of bandwidth. In this equation the bandwidth is calculated by a function $p(u,v)$, which is a function centered on the case located at (u,v) and based on the local population.

2.3 Summary

This chapter surveyed the literature concerning the probabilistic topic models (probabilistic Latent Semantic Analysis and the Latent Dirichlet Allocation model), their extensions and applications in the topic detection and tracking field, as well as different approaches to analyze data over time (Time Series) and space (geographic density maps).

As described in this chapter, the probabilistic topic models are very useful to classify a collection of documents in different topics. In brief, these models represent documents as a mixture of topics, where a topic is a probability distribution over words. This characteristic allows these models to relate documents with the same topic which contain different word distributions, as well as distinguish documents with polysemic words, of which these words could have a different meaning in different contexts.

By combining these models with additional components of the respective documents, it is possible to perform other tasks, such as topic trend discovery over time (as mentioned in Section 2.2.6), distinguish content of folksonomy resources (as mentioned in Section 2.2.4), discover the most interesting locations described in weblogs (as mentioned in Section 2.2.5) (Wang *et al.* (2007)), etc.

Finally, this chapter described different approaches to analyze data. This thesis soughts to analyze the topics trends of the collection of newswire documents over time using time series, and over space by using geographic density maps, of which can be generated by different kernel density estimators, such as inverse distance weighting and the Kriging interpolation method.

Chapter 3

Geo-temporal Topic Analysis

This chapter describes in detail the different tasks of the proposed approach mentioned briefly in Section 1.1, as well as the used software and its respective configuration in order to perform a geo-temporal topic analysis on the collection's associated topics.

3.1 Proposed Approach

The proposed approach is divided in three main tasks, namely (1) geo-temporal extraction, which consists on extracting the spatial and temporal information of each document, (2) topic assignment, which classifies the collection of newswire documents in different topics, (3) and finally by combining the outputs of these two tasks, graphics are generated which are later used to analyze the collection's topics over time and space. Figure 3.9 illustrates the pipeline process of the approach's prototype.

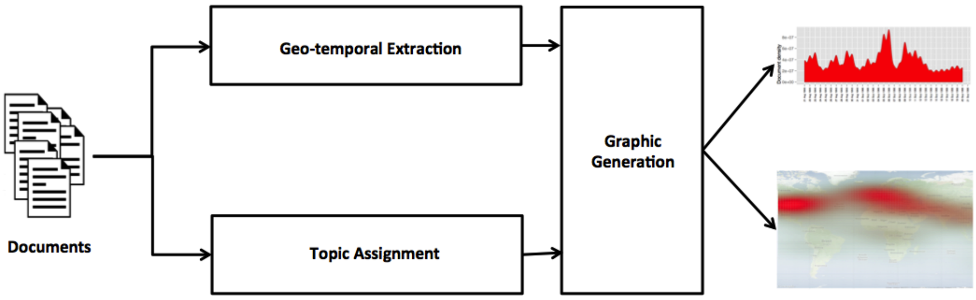

Figure 3.9: Proposed process pipeline of the approach's prototype.

3.1.1 Geo-temporal Extraction

Given a large collection of newswire documents, each document reports a happening in a specific time and geographic space.

The temporal extraction of each document is rather trivial, due to the fact that all newswire documents contain a publication date. Although the geographic extraction is not that linear as the first, due to this component is not usually discriminated in newswire documents. Therefore, for the proposed approach it was used the Yahoo! Placemaker[1] in order to indirectly extract this geographic component.

3.1.1.1 Yahoo! Placemaker

The Yahoo! Placemaker is a web service, which extracts the spatial information from a given document. In brief, the Yahoo! Placemaker uses natural language to disambiguate and extract geographic references within a document.

This web service is invoked via HTTP POST, which receives a document type (i.e. plain text) as input and returns a structure containing several kind of relevant geographic information (typically in XML format), namely a list of the detected geographic references, the geographic region which best describes the document, etc. Each of these elements is associated with a pair of geographic coordinates (latitude and longitude). Table 3.1 illustrates the most relevant elements for the task. For this thesis, we are particularly interested in only one of these components. This component is designated by *geographicScope* and it contains the reference of the smallest place which best describes the document and its corresponding geographic coordinates.

[1] http://developer.yahoo.com/geo/placemaker/

Element	Description
Administrative Scope	Element containing the smallest administrative place that best describes the document
Geographic Scope	Element containing the smallest place that best describes the document
Local Scope	Element containing the smallest named place associated with the document

Table 3.1: Most relevant elements of Yahoo Placemaker's response structure.

3.1.2 Topic Assignment

The topic assignment task is performed in two manners, namely (1) manually classifying the collection's documents using the already assigned topics of the collection, or (2) automatically classifying the collection's documents using a probabilistic topic model.

The automatic topic assignment is performed by using a software implementation of the Latent Dirichlet Allocation algorithm. In order to use this software implementation, the documents must be previously pre-processed. This pre-processing consists on representing the documents in document vectors, removing its stop words and stemming its remaining words.

Iniatially, it was used the R LDA Gibbs Sampler from the CRAN *lda* package (Blei *et al.* (2003), Chang (2011)). Due to hardware deficiency, later the C/C++ implementation also known as GibbsLDA++ [1] was adopted, given its lower memory consumption and more efficient processing. This software implementation comes in a form of a function, which receives several arguments as input, such as (1) a file containing the collection of documents, (2) a K integer representing the number of topics to consider in the model, (3) a number of iterations of Gibbs sampling to apply over the collection of which it was assigned 2000 iterations in the conducted experiments, (4) a scalar value Alpha which corresponds to the Dirichlet hyperparameter for topic proportions of which we consider its value to be 50/K topics (5) and finally the hyperparameter Beta for each entry of the block relations matrix, which we assign as default as 0.1.

The function returns several outputs, namely (1) a file containing the word-topic distributions ($p(word_w|topic_t)$), (2) a file containing the topic-document distributions ($p(topic_t|document_m)$), (3) a file containing the topic assignments for each word of the collection of documents, and (4) a file containing the most likely words for each topic.

3.1.3 Graphic Generation

Given the outputs of the two tasks mentioned in Sections 3.1.1 and 3.1.2, this final task will generate two types of graphics, namely time series in order to analyze the topic's temporal trends and geographic maps containing the topic's geographic distribution in order analyze each topic over space.

The time series are generated using a function designated by *geom_density* from the R package *ggplot2*, which displays a smooth density distribution of documents over time, thus helping us better analyze the different temporal trends (i.e. bursts) of the topic's temporal document distribution (Wickham (2009)). To be noted that the smoothing of the data is achieved by using a Gaussian Kernel Estimator.

[1] http://gibbslda.sourceforge.net/

The geographic maps containing each topic's geographic distribution are generated using the *kde2d* function from the the R package *MASS*, which is a two-dimensional kernel density estimator.

3.2 Summary

This chapter discussed the different tasks to generate the graphics in order to perform a geo-temporal topic analysis of a large collection of newswire documents. The tasks are namely (1) geo-temporal extraction, which extracts the document's temporal component, as well as the document's geographic component by using the Yahoo! Placemaker web service, (2) topic assignment of the collection's documents of which can assigned manualy or automatically using a Probabilistic Topic model, and finally (3) this chapter described how we generated the graphics in order to analyze the collection's topics over time and space.

Chapter 4

Experimental Evaluation

This chapter demonstrates through several experiments the effectiveness of the proposed approach to detect and track important events in a large collection of newswire documents over time and space.

Additionally, this chapter will describe the used dataset, the evaluation methodology, the conducted experiments, as well as discuss the achieved results.

4.1 Dataset

For the work's validation, it was conducted several experiments on the Reuters Corpus Volume 1 (RCV1), which is a collection of over 800,000 manually categorized newswire stories (Rose *et al.* (2002)). Each newswire story can be categorized over 55 topics, which are illustrated in Table 4.2.

This collection has a timespan of one year, which initiates on August of 1996 and terminates on August of 1997. Geographically, the newswire stories mainly occur in 3 regions, namely North America, Europe and Asia, as illustrated in Figure 4.10.

4.2 Evaluation Methodology

To evaluate the proposed approach, it was analyzed the document distribution trend of each topic over time, using time series (see in Section 2.2.7), and over space, using geographic maps displaying the geographic distribution of each topic generated with a Kernel Density Estimator (see Section 2.2.9).

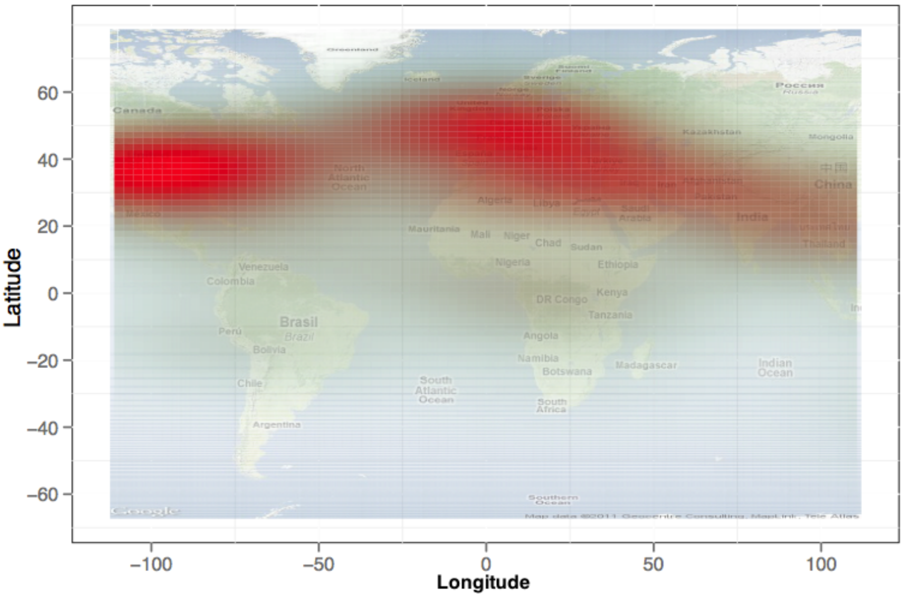

Figure 4.10: RCV1 Geographic document distribution.

The main objective with these topic analyses is to understand if indeed there is a correlation between topic patterns and the occurrence of important events.

In order to prove this correlation, the topic patterns were compared with a set of important events over time and space. By performing this comparison, it is possible to effectively verify if the time and space of the topic patterns actually correspond to the occurrence of important events.

This set of events was extracted from wikipedia, which consists of 187 events covering 25 of RCV1's manually assigned topics. Additionally the events have a very diverse temporal distribution, as illustrated in Figure 4.11. Also, its geographical distribution covers the same geographic regions as the RCV1 dataset as illustrated in Figure 4.12. In order to be able to compare these events with the topic patterns, each event was later associated with a RCV1 manual assigned topic, a generated Latent Dirichlet Allocation topic, a location of where the event took place, as well as its corresponding geographic coordinates which were automatically generated by Yahoo! Placemaker. Table 4.3 illustrates some examples of this set of important events.

For the temporal topic analysis, we used the mentioned techniques in Section 2.2.7 to analyze the different patterns of the temporal document distribution of each topic. By using these techniques, we will attempt to detect when an important event occurs in a time series, of which usually are displayed in a form of a burst.

As such, we developed an automatic evaluation method which measures, for each topic, the average minimum distance between an important event and its closest following detected burst. The bursts are automatically detected using a method designated by *msPeakSearch* from *msProcess*

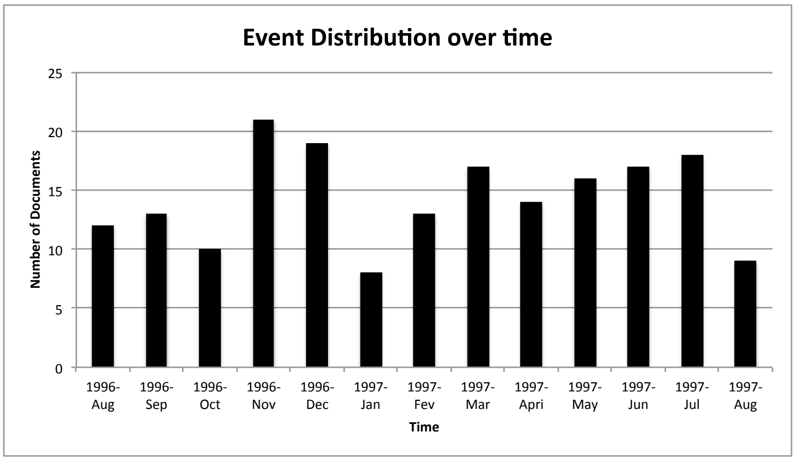

Figure 4.11: Temporal distribution of the set of important events.

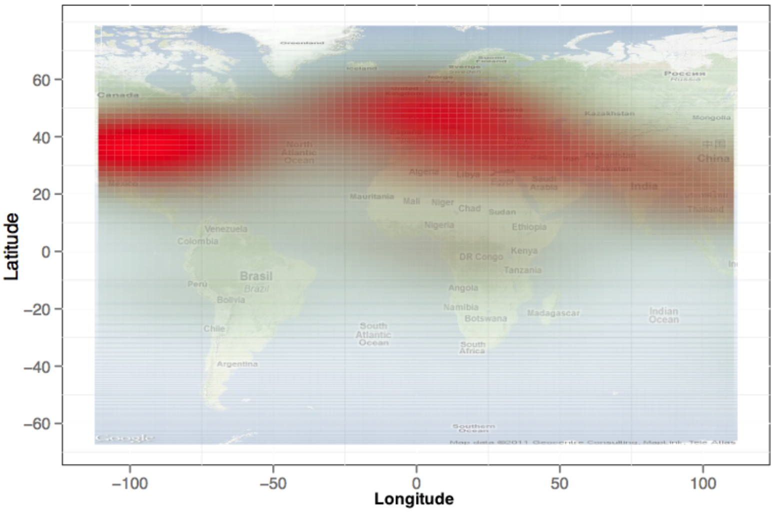

Figure 4.12: Geographic distribution of the set of important events.

R package, which seeks intensities that are higher than those in a local area and are higher than an estimated average background at the sites (Lixin Gong & Chen (2011)). Therefore by using this evaluation method, we can measure the effectiveness of a topic to detect and track important events over time.

For the geographic topic analysis, we used a kernel density estimator (see Section 2.2.9 for more detail) in order to display in which geographic regions a certain topic has a higher document density. As such, we will attempt to determine if indeed these high document densities actually correspond to a geographic region where an important event took place. In order to do so, for each important event, we will generate the geographic document density of the event's respective topic, of the analyzed documents that were published between 5 days before and after the respective event occurred, and observe if indeed the location of the geographic regions with the highest document density overlaps the geographic region where the important event took place.

4.3 Experiments

This section will describe the conducted experiments, as well discuss its achieved results.

4.3.1 Topic Assignment and Comparision

In the initial conducted experiments, it was analyzed two types of topics associated to the RCV1 collection, namely the already assigned RCV1's manually assigned topics (see Table 4.2) and the automatically generated topics from the Latent Dirichlet Allocation algorithm (LDA topics).
The Latent Dirichlet Allocation classified the collection in 55 different topics in attempt to recreate similar topics as the RCV1's manual assigned topics. This algorithm associated to each document of the collection a probabilistic topic distribution. According to this distribution, each document has a probability of belonging to each topic. Although for simplification reasons, we assigned to each document with the topic of which it had the highest probability to belong to.
In many cases the LDA topics were able to largely recreate the manual assigned topics, as illustrated in Table 4.4, where many of the LDA topics clustered a high percentage of the same documents as the RCV1's manually assigned topics.
Another proof of the sucess in discovering topics is the set of the most frequent words of each LDA topic, as illustrated in Table 4.5. This set of words could very well determine what the topic is about. For example the words of topic 7 (health, drug, medical, care, hospital) clearly represent the topic Health.

4.3.2 Temporal Topic Analysis

In the initial analyses, it was noticed that each topic displayed an unique trend over time, that could very well correspond to an event of some sort. For example as illustrated in Figure 4.13, the document density of the topic *Equity Markets* presented a periodic pattern which displayed a high document density on weekdays and a low document density on weekends. This phenomenon is explained by the fact that Equity Markets are closed on weekends. Additionally, its corresponding LDA topic, topic *24*, also displayed the same periodic pattern, as illustrated in Figure 4.14, proving once more the successful topic recreation of the Latent Dirichlet Allocation algorithm.

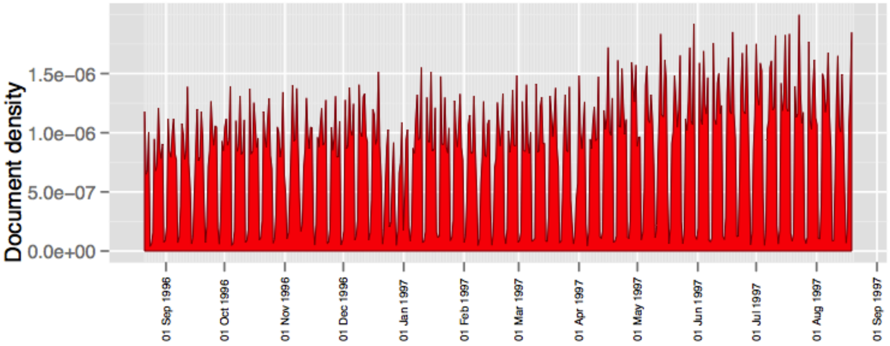

Figure 4.13: Document density over time for topic M11.

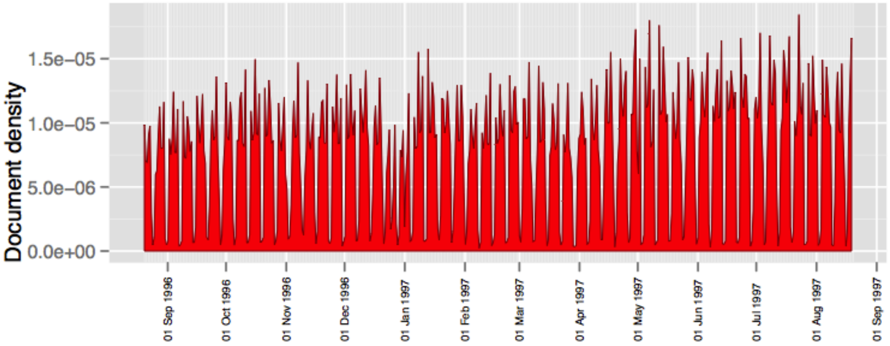

Figure 4.14: Document density over time for topic 24.

Another unique pattern that these temporal document distributions display are the *bursts* (see Section 2.2.7 for more detail). It is believed that there could be a correlation between this pattern and the occurrence of an important event, and therefore this pattern should be analyzed in more detail. As illustrated in Figure 4.15 and 4.16, the temporal document distribution of Topic *Disasters and Accidents* and its corresponding LDA topic *51*, both contain a serie of bursts. After manually analyzing all the documents related to the dates of each burst, its was detected that indeed the documents of the respective burst mainly described about an important event. For example both of the topics displayed two major bursts, which in fact correspond to two important events, namelly on the 5th September 1996 Hurricane Fran arrived South Carolina and subsequently on the 9th of September 1996 this same hurricane dissipated.

To further prove this correlation between the occurrence of bursts and important events, the timestamps of the bursts were compared with the timestamps of set of important events. In order to perform this comparison, it was generated for each topic a time series containing the temporal document distribution of the respective topic, an indicator of when an important event occurred (black vertical line) and an indicator of a detected burst (blue vertical line). For example as

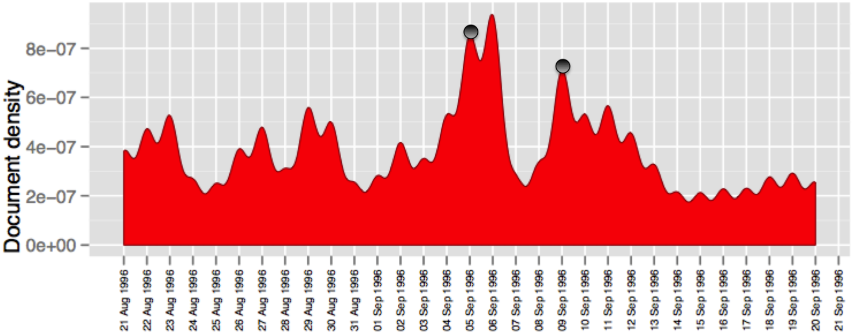

Figure 4.15: Document density over time for topic GDIS.

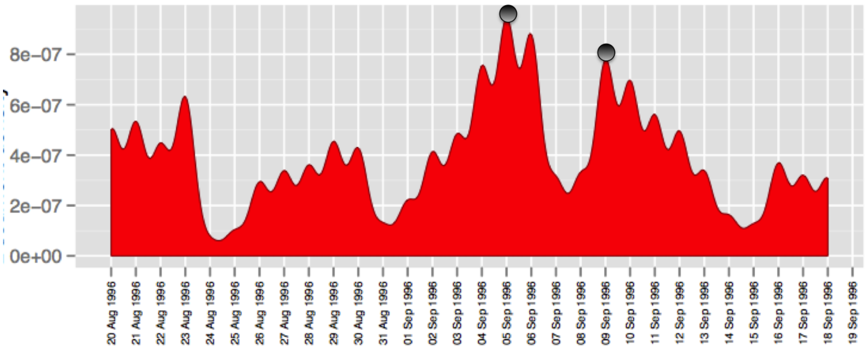

Figure 4.16: Document density over time for topic 51.

illustrated in Figure 4.17, we can see that most of the important events of the topic *Elections* correspond very closely to the same timestamps as the detected bursts. However, it was also noted that the number of detected bursts is significatly greater then the number of important events.This is justified by the fact that these additional detected bursts correspond to less important events which do not have enough importance to belong to the set of important events.

Additionally, we used the described evaluation method in Section 4.2 in order to evaluate the topic's effectiveness to detect and track important events over time, of which Table 4.6 demonstrates the results of the RCV1's topics. In general the temporal document distributions displayed significantly more bursts then the compared important events, however the average minimum distances between an event and its following closest burst of most topics are rather small.
In Table 4.7, we can see the results of the LDA topics, which display a better average minimum distances then the RCV1's. This can be justified by the fact that the LDA topic temporal document distributions displayed substantially more bursts then the RCV1 manual assigned topics.

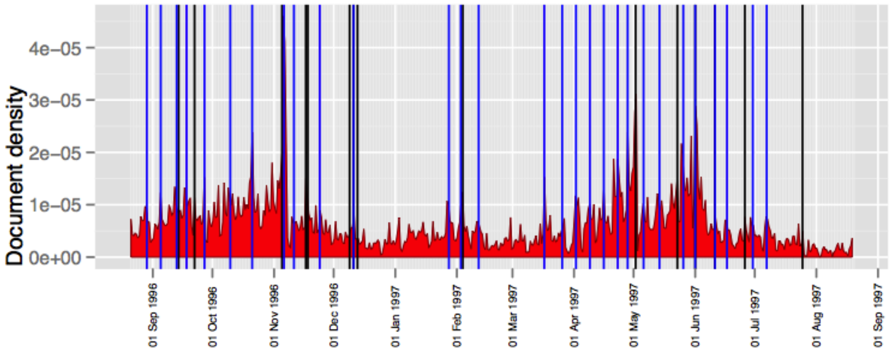

Figure 4.17: Document density over time for topic GVOTE.

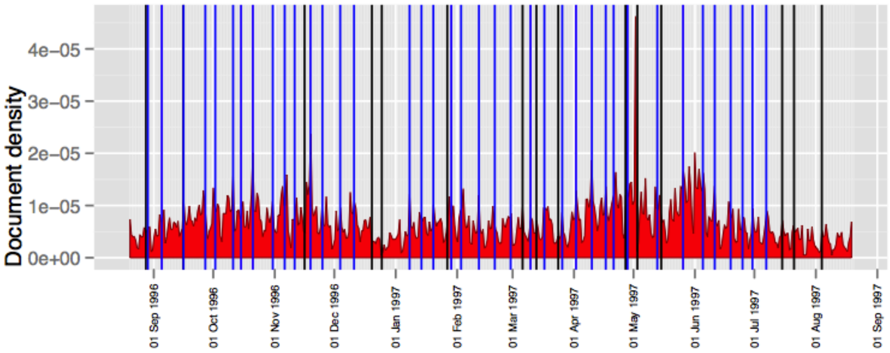

Figure 4.18: Document density over time for topic 25.

4.3.3 Geographic Topic Analysis

The geographic topic analysis is a rather easier analysis to visualy evaluate. By using the previously described evaluation method in Section 4.2, it was determined if indeed there is a correlaction between the geographic regions with the highest document density of a certain topic and the location of where an important event took place.

Figure 4.19 shows a clear example that there is in fact a strong correlation between the two. As observed in this Figure, the geographic regions where the topic *War, Civil War* has a higher document density, actually overlaps with the same geographic region as the important event, which is represented with a black point (Iraq disarmament crisis: Iraqi forces launch an offensive into the northern No-Fly Zone and capture Arbil). By geographically analyzing the corresponding LDA topic of topic *War, Civil War*, it was also possible to detect the same event geographically. As illustrated in Figure 4.20, the geographic document distribution of topic *1* displays a high document density in the same geographic region as the important event. However, it was also noted that

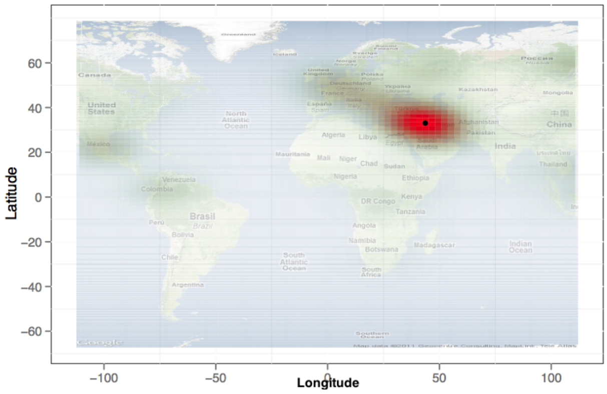

Figure 4.19: Document density over space for topic GVIO.

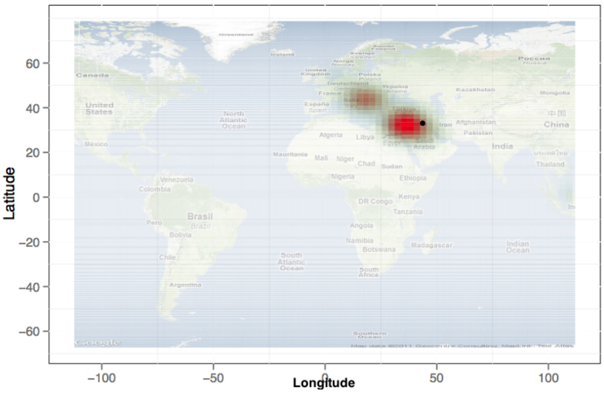

Figure 4.20: Document density over space for topic 1.

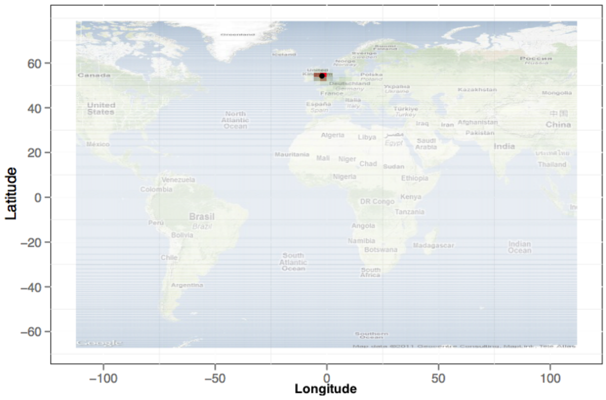

Figure 4.21: Document density over space for topic GVOTE.

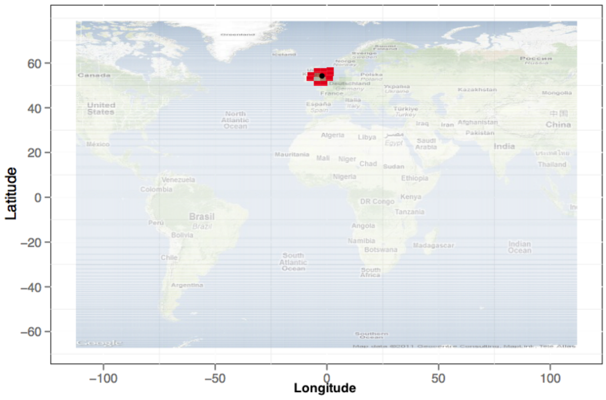

Figure 4.22: Document density over space for topic 11.

the geographic document distribution of this LDA topic is less accurate to geographically pinpoint an important event then its corresponding manual assigned topic.

4.4 Summary

This chapter presented and discussed the conducted experiments as well as its achieved results. On the temporal topic analysis, it was determined that indeed there is a strong correlation between the temporal burst pattern and an important event.

The geographic topic analysis was proven as well as to be effective to detect important events over space, by exploring the geographic regions with a high topic density of which usually corresponds to a geographic region where an important event took place.

Both of these correlations were proven with the proposed evaluation method.

Finally, it was proven that the generated topics of the Latent Dirichlet Allocation algorithm are suitable to detect and track important events over time and space.

Topics	Description	Nr. of Documents
C11	STRATEGY.PLANS	24325
C12	LEGAL/JUDICIAL	11944
C13	REGULATION/POLICY	37410
C14	SHARE LISTINGS	7410
C15	PERFORMANCE	151784
C16	INSOLVENCY/LIQUIDITY	1920
C17	FUNDING/CAPITAL	42155
C18	OWNERSHIP CHANGES	52817
C21	PRODUCTION/SERVICES	25403
C22	NEW PRODUCTS/SERVICES	6119
C23	RESEARCH/DEVELOPMENT	2625
C24	CAPACITY/FACILITIES	32153
C31	MARKETS/MARKETING	40509
C32	ADVERTISING/PROMOTION	2084
C33	CONTRACTS/ORDERS	15332
C34	MONOPOLIES/COMPETITION	4835
C41	MANAGEMENT	11355
C42	LABOUR	11878
E11	ECONOMIC PERFORMANCE	8568
E12	MONETARY/ECONOMIC	27100
E13	INFLATION/PRICES	6603
E14	CONSUMER FINANCE	2177
E21	GOVERNMENT FINANCE	43130
E31	OUTPUT/CAPACITY	2415
E41	EMPLOYMENT/LABOUR	17035
E51	TRADE/RESERVES	21280
E61	HOUSING STARTS	391
E71	LEADING INDICATORS	5270
G15	EUROPEAN COMMUNITY	20658
GCRIM	CRIME & LAW ENFORCEMENT	32219
GDEF	DEFENCE	8842
GDIP	INTERNATIONAL RELATIONS	37739
GDIS	DISASTERS AND ACCIDENTS	8657
GENT	ARTS & CULTURE & ENTERTAINMENT	3801
GENV	ENVIRONMENT AND NATURAL WORLD	6261
GFAS	FASHION	313
GHEA	HEALTH	6030
GJOB	LABOUR ISSUES	17241
GMIL	MILLENNIUM ISSUES	5
GOBIT	OBITUARIES	844
GODD	HUMAN INTEREST	2802
GPOL	DOMESTIC POLITICS	56878
GPRO	BIOGRAPHIES & PERSONALITIES & PEOPLE	5498
GREL	RELIGION	2849
GSCI	SCIENCE & TECHNOLOGY	2410
GSPO	SPORTS	35316
GTOUR	TRAVEL & TOURISM	680
GVIO	WAR & CIVIL WAR	32615
GVOTE	ELECTIONS	11532
GWEA	WEATHER	3878
GWELF	WELFARE AND SOCIAL SERVICES	1869
M11	EQUITY MARKETS	48700
M12	BOND MARKETS	26036
M13	MONEY MARKETS	53633
M14	COMMODITY MARKETS	85446

Table 4.2: RCV1 list of topics and respective document distributions.

Date	RCV1 Topic	LDA Topic	Geographic Region	Latitude	Longitude	Event Description
20-Aug-1996	GCRIM	6	Korean Peninsula	38.2218	126.994	A thousands-large protest in Seoul calling for reunification with North Korea is broken up by riot police.
21-Aug-1996	GCRIM	6	South Africa	-28.4793	24.6799	Former president of South Africa F. W. de Klerk makes an official policy for crimes committed under Apartheid to the Truth and Reconciliation Commission in Cape Town.
31-Aug-1996	GSPO	16	Manhattan, KS, US	39.1788	-96.5618	The Big 12 Conference is inaugurated with a football game between Kansas State University and Texas Tech University in Manhattan, Kansas.
3-Sep-1996	GVIO	1	Iraq	33.2405	43.6898	The U.S. launches Operation Desert Strike against Iraq in reaction to the attack on Arbil.
14-Sep-1996	GVOTE	55	Bosnia and Herzegovina	43.9201	17.677	Alija Izetbegoviƒá is elected president of Bosnia and Herzegovina in the country's first election since the Bosnian War.
5-Feb-1997	GODD	19	Switzerland	46.8132	8.22395	The so-called ""Big Three"" banks in Switzerland announced the creation of a $71 million fund to aid Holocaust survivors and their families.
10-Feb-1997	GCRIM	4	United States	37.1679	-95.845	The United States Army suspends Gene C. McKinney Sergeant Major of the Army its top-ranking enlisted soldier after hearing allegations of sexual misconduct.
6-Jun-1997	GCRIM	10	Lacey, NJ, US	39.8595	-74.2067	In Lacey Township New Jersey high school senior Melissa Drexler kills her newborn baby in a toilet.
13-Jun-1997	GCRIM	4	Oklahoma City, OK, US	35.472	-97.5203	A jury sentences Timothy McVeigh to death for his part in the 1995 Oklahoma City bombing.
8-Jul-1997	GDIP	55	European Union	50.2528	5.56973	NATO invites the Czech Republic, Hungary and Poland to join the alliance in 1999.
10-Jul-1997	GSCI	19	London, England GB	51.5063	-0.12714	In London, scientists report their DNA analysis findings from a Neanderthal skeleton, which support the out of Africa theory of human evolution placing an "African Eve" at 100.000 to 200.000 years ago.
10-Jul-1997	GCRIM	23	Ermua, Basque Country, ES	43.1792	-2.49738	Miguel Ångel Blanco is kidnapped in Ermua, Spain and murdered by the ETA.
1-Aug-1997	GPRO	37	Boston, MA, US	42.3586	-71.0567	Steve Jobs returns to Apple Computer, Inc at Macworld in Boston.
2-Aug-1997	GDIS	10	New South Wales AU	-32.831	147.319	Australian ski instructor Stuart Diver is rescued as the sole survivor from the Thredbo landslide in New South Wales in which 18 die.
6-Aug-1997	GDIS	31	Guam	13.4465	144.787	Korean Air Flight 801 crash lands west of Guam International Airport, resulting in the deaths of 228 people.

Table 4.3: Set of Important Events.

RCV1 Topic	Description	LDA Topic	RCV1 # documents	LDA # documents	Percentage of LDA Documer
C11	Strategy Plans	31	24325	2778	11
C12	Legal/Judicial	23	11944	5056	42
C13	Regulation/Policy	7	37410	2963	8
C14	Share Listings	13	7410	2442	33
C15	Performance	43	151782	40983	27
C16	Insolvency/Liquidity	13	1920	448	23
C17	Funding/Capital	13	42154	12242	29
C18	Ownership Changes	13	52816	14655	28
C21	Production/Services	20	25403	4875	19
C22	New Products/Services	6	6119	2123	35
C23	Research/Development	7	2625	1559	59
C24	Capacity/Facilities	31	32153	8390	26
C31	Markets/Marketing	20	40509	3872	10
C32	Advertising/Promotion	6	2083	793	38
C33	Contracts/Orders	6	15332	2521	16
C34	Monopolies/Competition	2	4835	1037	21
C41	Management	6	11355	4556	40
C42	Labour	29	11878	5589	47
E11	Economic Performance	21	8568	3551	41
E12	Monetary/Economic	55	27100	6388	24
E13	Inflation/Prices	37	6603	2500	38
E14	Consumer Finance	37	2177	784	36
E21	Goverment Finance	14	43130	17889	41
E31	Output/Capacity	37	2415	1443	60
E41	Employment/Labour	29	17035	5868	34
E51	Trade/Reserves	40	21280	4553	21
E61	Housing Starts	37	391	163	42
E71	Leading Indicators	53	5270	4568	87
G15	European Community	55	20658	7064	34
GCRIM	Crime& Law Enforcement	23	32219	13724	43
GDEF	Defense	45	8842	1813	21
GDIP	Internacional Relations	40	37739	8630	23
GDIS	Disasters And Accidents	51	8657	5651	65
GENT	Arts& Culture& Entertainment	49	3801	2720	72
GENV	Environment And Natural World	51	6261	1674	27
GFAS	Fashion	49	313	239	76
GHEA	Health	7	6030	3866	64
GJOB	Labour Issues	29	17241	6146	36
GMIL	Millennium Issues	6	5	1	20
GOBIT	Obituaries	49	844	481	57
GODD	Human Interest	49	2802	1526	54
GPOL	Domestic Politics	25	56878	12978	23
GPRO	Biographies& Personalities & People	49	5498	2113	38
GREL	Religion	49	2849	1021	36
GSCI	Science And Technology	28	2410	956	40
GSPO	Sports	52	35314	15023	43
GTOUR	Travel And Tourism	49	680	118	17
GVIO	War & Civil War	10	32615	9077	28
GVOTE	Elections	25	11532	6379	55
GWEA	Weather	51	3878	1666	43
GWELF	Welfare & Social Services	19	1869	674	36
M11	Equity Markets	24	48700	31979	66
M12	Bond Markets	3	26036	5932	23
M13	Money Markets	9	53633	26879	50
M14	Commodity Markets	20	85446	19674	23

Table 4.4: RCV1's Topic Distribution.

Topic	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Word 11	Word 12	Word 13	Word 14	Word 15	Word 16	Word 17	Word 18	Word 19	Word 20
1	israel	israeli	iraq	peace	al	east	minister	netanyahu	iran	saudi	west	iraqi	president	security	arafat	prime	united	lebanon	official	arabia
2	week	pakistan	ago	test	day	chicago	light	top	demand	kabul	total	pakistani	year	west	bhutto	afghanistan	match	movement	runs	early
3	business	market	investment	years	world	major	global	international	small	large	future	largest	part	make	billion	big	capital	industrial	services	making
4	week	pakistan	ago	test	day	chicago	light	top	demand	kabul	total	pakistani	year	west	bhutto	afghanistan	match	movement	runs	early
5	british	london	uk	ireland	northern	major	made	building	david	announced	investment	royal	business	scotland	international	company	order	year	tony	national
6	market	percent	closed	stock	close	street	day	share	gains	week	blue	trade	dow	high	early	average	industrial	jones	issues	news
7	health	food	drug	medical	care	study	hospital	found	british	people	heart	university	disease	caused	drugs	treatment	mother	human	said	research
8	million	city	state	school	general	texas	california	hotel	service	property	florida	building	chicago	san	total	board	jersey	department	los	information
9	market	day	trade	close	closed	early	week	high	support	london	demand	gains	morning	bought	month	light	earlier	chicago	held	mid
10	military	army	government	forces	troops	war	rebels	force	zaire	refugees	fighting	rebel	people	capital	president	peace	soldiers	mobutu	aid	country
11	government	minister	state	stories	prime	leading	official	national	cabinet	newspapers	country	bulgaria	president	trade	news	transport	office	owned	general	board
12	union	italy	countries	government	budget	france	minister	join	spain	treaty	french	policy	planned	country	member	made	time	make	entry	launch
13	share	loss	months	total	note	average	thousands	general	fully	results	research	charges	pro	stock	dollars	chicago	company	corporation	services	investment
14	budget	government	house	bill	state	clinton	senate	billion	reform	federal	congress	republican	cuts	president	pay	law	proposed	year	years	system
15	bank	billion	banks	fund	investment	capital	international	term	thai	thailand	system	asset	total	property	country	government	including	reserve	worth	largest
16	romania	news	east	billion	year	planned	earlier	business	months	pay	made	early	report	office	top	conference	mid	leading	part	chemical
17	canada	government	city	de	colombia	president	american	peru	country	cuba	venezuela	hostages	lima	release	drug	states	residence	winds	japanese	storm
18	police	people	killed	city	fire	town	security	km	miles	area	bomb	state	dead	attack	injured	shot	southern	killing	official	protest
19	swedish	research	norway	months	system	general	hold	values	issues	section	started	recommendation	report	direction	source	order	leading	call	ice	largest
20	percent	year	billion	month	total	months	trade	average	earlier	figure	demand	ago	industrial	orders	released	department	services	national	report	office
21	year	million	market	official	morgan	issues	billion	treasury	international	bank	years	global	term	senior	week	note	stanley	london	future	launch
22	percent	policy	year	report	federal	reserve	bank	government	jobs	months	chief	market	average	high	pace	signs	released	greenspan	demand	term
23	court	case	charges	law	trial	federal	legal	found	justice	state	years	office	allegations	death	police	prison	supreme	illegal	department	arrested
24	stock	million	share	market	closed	percent	trade	worth	day	close	company	total	high	hands	capital	american	orders	big	small	bought
25	party	government	election	minister	opposition	prime	elections	leader	president	national	parties	democratic	support	country	general	people	leaders	congress	socialist	presidential
26	million	tonne	week	total	bought	chicago	world	red	sole	crops	board	brazil	average	year	northern	demand	area	report	official	dollars
27	australian	australia	zealand	national	howard	year	south	government	wales	pacific	international	morning	announced	week	major	news	federal	david	made	high
28	services	computer	service	company	internet	information	business	system	news	digital	electronic	international	microsoft	equipment	intel	launch	ibm	apple	radio	including
29	time	make	people	big	made	years	difficult	making	decision	work	future	great	weeks	major	months	news	days	forced	job	makes
30	company	million	billion	greek	signed	owned	total	day	area	state	agreement	north	percent	sea	work	west	begin	island	part	canada
31	india	indian	bank	industries	results	corporation	state	industrial	capital	global	call	note	government	time	investment	pradesh	andhra	international	information	market
32	united	states	trade	countries	agreement	world	nations	president	secretary	minister	members	conference	international	leaders	issues	official	human	peace	called	progress
33	workers	strike	union	nuclear	work	hours	company	pay	day	general	service	jobs	shut	early	strikes	industrial	week	light	agreement	part
34	china	kong	hong	chinese	philippines	property	pacific	people	official	news	thailand	handover	land	international	british	island	rule	tung	years	future
35	chief	company	chairman	president	board	business	vice	named	senior	services	announced	appointed	office	general	international	american	top	david	michael	job
36	french	de	france	billion	brazil	percent	state	brazilian	paulo	spanish	spain	la	jean	belgium	owned	government	part	chairman	announced	capital
37	food	ban	department	environmental	states	world	land	year	years	united	report	make	made	national	chemical	small	including	country	called	worth
38	pay	call	entry	bank	london	announced	orders	general	total	force	time	law	home	rule	capital	morgan	school	english	stanley	note
39	market	high	term	week	support	demand	weeks	month	early	hold	remains	large	big	time	major	mid	research	top	months	days
40	uk	trade	industrial	week	total	note	food	average	orders	base	budget	leading	home	official	services	building	business	capital	job	source
41	russia	russian	president	nato	turkey	news	minister	security	state	alliance	prime	year	republic	operation	signed	heart	military	month	poland	made
42	fuel	tonne	demand	west	days	sea	week	coast	general	day	light	east	london	north	major	high	texas	source	km	discharge
43	million	year	company	share	results	billion	percent	business	ago	earlier	report	store	ending	loss	call	charges	stock	including	gains	announced
44	company	stock	million	share	merger	percent	capital	agreement	owned	board	proposed	announced	held	bought	largest	investment	planned	distribution	class	pay
45	commission	council	report	decision	union	member	conference	system	national	law	policy	proposed	states	aid	trade	case	legal	information	state	order
46	time	year	years	made	pay	future	close	term	operation	ago	successful	company	land	strategic	history	chief	broken	alliance	million	founded
47	world	match	league	win	won	game	italy	home	top	champion	france	spain	victory	year	time	champions	united	english	winning	women
48	copper	year	world	london	mine	tonne	demand	month	months	official	base	close	high	largest	russian	earlier	major	general	market	source
49	years	world	year	people	family	day	women	born	home	life	time	great	war	work	history	age	american	mother	star	son
50	percent	bank	day	month	week	market	treasury	average	days	billion	banks	bill	reserve	term	call	worth	national	weeks	demand	total
51	air	airlines	airport	aircraft	service	american	traffic	space	boeing	flight	international	passenger	transport	plane	airways	services	crew	route	crash	fuel
52	game	san	win	won	time	home	chicago	made	runs	league	victory	year	shot	boston	los	national	angeles	paul	david	winning
53	million	year	ago	loss	poland	months	note	equipment	makes	chemical	major	leading	industries	chain	large	railway	electronic	comprehensive	industrial	store
54	czech	hungary	republic	bosnia	international	bosnian	state	war	year	national	serbian	country	results	nato	news	zagreb	made	held	part	capital
55	dollar	japanese	bank	market	trade	dollars	morning	close	early	closed	month	demand	earlier	high	opened	official	major	senior	support	large

Table 4.5: 20 most frequent words of each generated LDA Topic.

Topic	Minimum Average Distance between Event and Burst (Days)	Nr. of Events	Nr. of Bursts
C11	5	1	49
C12	8	1	48
C14	2	1	45
C15	0.5	2	42
C18	6.5	4	48
C23	0	1	2
GCRIM	3.38	29	48
GDIP	3.67	9	44
GDIS	4.91	32	38
GENT	13.9	10	13
GENV	0	1	46
GHEA	2	1	45
GJOB	4	4	46
GOBIT	23.34	9	3
GODD	51.35	12	3
GPOL	3.83	6	47
GPRO	6.67	9	31
GREL	0	1	6
GSCI	58.41	12	5
GSPO	0.67	3	45
GTOUR	0	3	3
GVIO	2.84	19	45
GVOTE	6	15	29
GWELF	16	1	37
M11	1	1	48

Average of average distances	8.31

Table 4.6: Average Minimum Distances between the topic's important events and its detected bursts of RCV1's topics.

Topic	Minimum Average Distance between Event and Burst (Days)	Nr. of Events	Nr. of Bursts
1	3.86	7	45
2	2.67	3	46
3	7.67	3	46
4	2.65	14	46
6	6.6	5	48
7	0	1	49
8	0.5	2	46
10	3.29	28	42
11	6.67	3	44
12	6	1	15
14	2.38	8	41
16	5.17	6	40
18	4.67	3	48
19	2.86	14	50
23	3.08	12	49
24	2.78	9	50
25	4.69	16	39
28	4	1	46
30	4.5	4	44
31	3	22	50
32	4	1	49
33	6.521	2	46
34	4	6	38
35	7	1	47
37	6.75	4	49
39	3.33	3	39
40	3	1	44
43	0	1	33
46	1	1	50
48	2	1	13
51	0	1	46
55	2.67	3	42

Average of average distances	3.66

Table 4.7: Average Minimum Distances between the topic's important events and its detected bursts of the LDA Topics.

Chapter 5

Conclusions

In this thesis, it was sought to explore the geo-temporal topic analysis of newswire documents. The topic assignment of the collection's documents was performed in two manners, namely by manually classifying the documents with a pre-determined list of topics, or automatically classifying the documents using a probabilistic topic model (Latent Dirichlet Allocation).

By examining the temporal and spatial trends of both of these type of topics, it was proven that these trends can effectively display relevant spatial and temporal information regarding events, by analyzing the discussed topics over time by generating time series, over space by generating geographic maps displaying the geographic distribution of each topic that was produced by a Kernel Density Estimator. From this analysis derived the proposed novel approach to detect and track important events over time and space.

This approach is based on the assumption that there is a correlation between geo-temporal topic patterns and occurrence of important events. In order to prove this assumption, it was proposed and used an evaluation method which evaluates the topic effectiveness to detect and track important events over time and space. By using the proposed evaluation method, it was proven that the proposed approach can effectivelly detect and track important events over time and space.

5.1 Contributions

The main contributions of this work are the following:

- A prototype was developed which followed the process of the proposed approach. This prototype generated graphics that were later used to perform a temporal topic analysis, in the case of the time series, and a geographic topic analysis by using geographic maps displaying the geographic distribution of each topic.

- After the geo-temporal topic analysis of the given large collection of newswire documents, it was determined that indeed there is a correlation between the geo-temporal patterns of the topics and when and where an important event took place.
- An evaluation method was proposed and used to validate our correlation assumption, as well as evaluate the effectiveness of each topic to detect and track important events over time and space.
- Using the proposed evaluation method, it was proven that the generated topics of the Latent Dirichlet Allocation algorithm were suitable for the task. Additionally, the achieved results from the evaluation method were used to compare the correspondence between the automatically generated topics and the manual assigned topics, of which the automatically assigned topics are more suitable to detect and track important events over time. Although the manual assigned topics achieved better results in the geographic detection and tracking.

5.2 Future work

Despite the achieved results, there are still some challenges for future research in the area of topic analysis in newswire documents that were not addressed due to lack of time. In this section I suggest some relevant opportunities for follow-up work:

- Integration of the Geofolk model in the topic analysis of newswire documents (Sizov (2010)). This model could be applied to answer several questions regarding the geographic distribution of topics, such as ”Which are the most relevant topics of each geographic region?”, ”Given a newswire document, which geographic region would be more concerned with its describing topic?”, etc.
- Integration of the proposed adaptation of Latent Dirichlet Allocation algorithm to analyze the relation between documents in the topic analysis of newswire documents (Wahabzada *et al.* (2010)). This model could be used to understand the potential interrelation between topics over time and space.

Bibliography

ALLAN, J., ed. (2002). *Topic detection and tracking: event-based information organization*. Kluwer Academic Publishers, Norwell, MA, USA.

BLEI, D.M., NG, A.Y. & JORDAN, M.I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*.

CARLOS, H., SHI, X., SARGENT, J., TANSKI, S. & BERKE, E. (2010). Density estimation and adaptive bandwidths: A primer for public health practitioners. *International Journal of Health Geographics*.

CHANG, J. (2011). *Collapsed Gibbs sampling methods for topic models*.

DEERWESTER, S., DUMAIS, S.T., FURNAS, G.W., LANDAUER, T.K. & HARSHMAN, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*.

DEMIRHAN, M., ÖZPINAR, A. & ÖZDAMAR, L. (2003). Performance evaluation of spatial interpolation methods in the presence of noise. *International Journal of Remote Sensing*.

DEMPSTER, A.P., LAIRD, N.M. & RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*.

DOBSON, J., COLEMAN, P., DURFEE, R. & WORLEY, B. (2000). Landscan: a global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing*.

GILKS, W. & SPIEGELHALTER, D. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC.

HOFMANN, T. (1999). Probabilistic latent semantic analysis. *In Proceedings of the 15th Conference on Uncertainty in AI*.

HURST, M. (2006). Temporal text mining. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*.

KOHONEN, T., SCHROEDER, M.R. & HUANG, T.S., eds. (2001). *Self-Organizing Maps*. Springer-Verlag New York, Inc., 3rd edn.

LI, L. & REVESZ, P. (2004). Interpolation methods for spatio-temporal geographic data. *Computers, Environment and Urban Systems*.

LIXIN GONG, W.C. & CHEN, Y.A. (2011). *Protein Mass Spectra Processing*.

MANNING, C.D., RAGHAVAN, P. & SCHÜTZE, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

MEI, Q., LIU, C., SU, H. & ZHAI, C. (2006). A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th international conference on World Wide Web*, ACM.

ROSE, T., STEVENSON, M. & WHITEHEAD, M. (2002). The reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In *In Proceedings of the 3rd International Conference on Language Resources and Evaluation*.

RZESZUTEK, R., ANDROUTSOS & KYAN, M. (2010). Self-organizing maps for topic trend discovery. *IEEE signal Processing Letters*.

SIZOV, S. (2010). Geofolk: latent spatial semantics in web 2.0 social media. In *Proceedings of the third ACM international conference on Web search and data mining*, ACM.

STEYVERS, M. & GRIFFITHS, T. (2007). *Probabilistic Topic Models*. Lawrence Erlbaum Associates.

WAHABZADA, M., XU, Z. & KERSTING, K. (2010). Topic models conditioned on relations. In *Proceedings of the European Conference on Machine Learning*.

WANG, C., WANG, J., XIE, X. & MA, W.Y. (2007). Mining geographic knowledge using location aware topic model. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*.

WICKHAM, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York.

W.REESE, K. & SHAFTO, P. (????). Towards a temporal latent dirichlet allocation model. Tech. rep., University of Louisville.