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

Erik Michael Leal Wennberg
erik.wennberg@ist.utl.pt

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

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. This paper is organized as follows: Section 2 presents some fundamental concepts. Section 3 presents the related work. Section 4 describes in detail the different tasks involved in the proposed approach, as well as the software used and its respective configuration. Section 5 describes the evaluation methodology and discusses the obtained results of the proposed approach, and finally Section 6 presents the achieved conclusions and future work which were derived from this thesis.

2. FUNDAMENTAL CONCEPTS

This thesis is framed in the Topic Detection and Tracking (TDT), which is an area of information retrieval that 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) [1]. 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 noticed 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 keywords, 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 [10]. 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 [10]. The similarity of documents represented in the bag-of-words model can be computed through vector matching operations such as the cosine similarity [10].
A Topic Distribution model represents documents as 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.
Two of the most widely used topic distribution models are the probabilistic Latent Semantic Analysis (pLSA) model and the Latent Dirichlet Allocation (LDA) model [13]. 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 [10].
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.

3. RELATED WORK

This section presents some of the most relevant related works for this thesis.

3.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic model for collections of discrete data, such as texted documents [2].
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.
LDA model, illustrated in Figure 1, 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 × V matrix β where $\beta_{ij} = p(w^j = 1|z^i = 1)$.
The LDA generative process can be sumarized in the following expression:

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

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 $z_n^i = 1$.

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.
Although there is a main inferential problem that must be solved in order to use the LDA model, the hidden variables

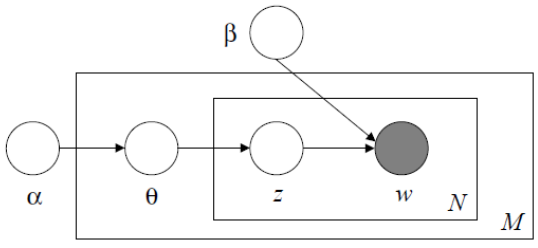

Figure 1: Graphical representation of LDA model using plate notation [2].

distribution (θ_d, z_n, β) must be previously computed, and this distribution is given by a document w:

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}. \quad (2)$$

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 [5] or Markov Chain Monte Carlo (MCMC) [7] methods such as Gibbs sampling [2].

3.1.1 The Gibbs Sampling Procedure for LDA

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:

$$P(z_i = j|z_{-i}, w_i, d_i, .) = \frac{C^{WT}_{w_ij} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_ij} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_it} + T\alpha} \quad (3)$$

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 (3) 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:

$$\theta^{(d)}_j = \frac{C^{DT}_{dj} + \alpha}{\sum_{k=1}^{T} C^{DT}_{dk} + T\alpha} \quad (4)$$

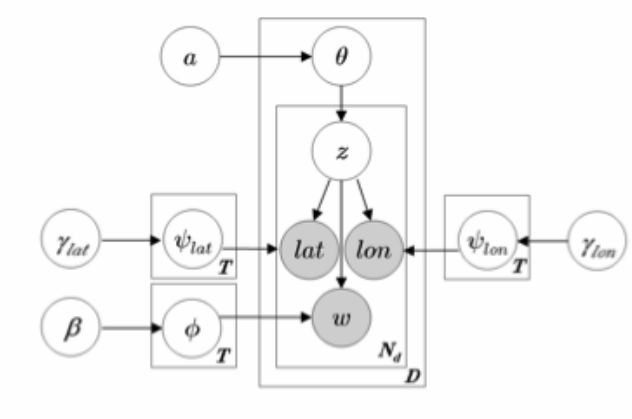

Figure 2: Graphical representation of the Geofolk model using plate notation [12].

3.2 The Geofolk Model

The Geofolk model is an extension of the LDA model for topic discovery using spatial information and word co-occurrences [12]. 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.

The generative process behind the Geofolk model for resources annotated with coordinates, illustrated in Figure 2, starts by executing LDA's steps as previously described in Section 3.1. In parallel, the topic generates two coordinates simultaneously, lat_{d_i} and lon_{d_i} from two topic-specific Gaussian distributions, ψ^{lat}_z and ψ^{lon}_z respectively. It is assumed that the Gaussian parameters μ^{lat}_z and μ^{lon}_z (i.e., the means for topic-specfic latitude and longitude) to compute ψ^{lat}_z and ψ^{lon}_z 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,

[1] `www.flickr.com`

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.

3.2.1 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.

3.2.2 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 3.
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.

3.3 Temporal Text Mining

Text mining on newswire 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 [8].
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 $t(i)/t_{bg}(i)$.

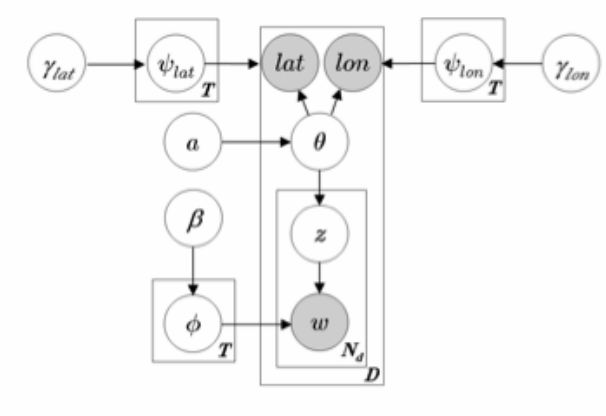

Figure 3: Graphical representation of the alternative Geofolk model for suggesting query locations using plate notation [12].

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:

$$\frac{t(q)}{(\sum_{i=p}^{q} t(i))/(1+q-p)} \tag{5}$$

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.

3.4 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 [3]. 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:

$$\lambda(x,y) = \frac{n}{|A|} \tag{6}$$

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:

$$f(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K(\frac{d_i}{h}) \tag{7}$$

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 [6]. 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:

$$f(x,y) = \sum_{i=1}^{n} K(\frac{d_i}{p(u,v)}) \tag{8}$$

The main diference between this equation and Equation 7 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.

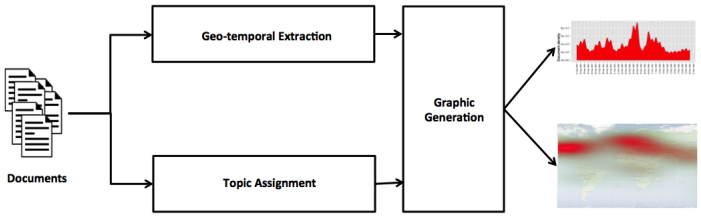

Figure 4: Proposed process pipeline of the prototype.

4. PROPOSED APPROACH

For this thesis, it was proposed a new approach to detect and track important events over time and space, by performing a geo-temporal topic analysis on a large collection of newswire documents. In order to analyze the collection's topics, it was developed a prototype that receives a collection of newswire documents and in turn generates graphics of which later which are later used to analyze the collection's topics over time and space.
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 4 illustrates the pipeline process of the proposed approach prototype.

4.1 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[2] in order to indirectly extract this geographic component.

4.1.1 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 1 illustrates the most relevant elements for the task.
For this MSc thesis, we are particularly interested in only one of these components. This component is designated by

[2]`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 1: Most relevant elements of Yahoo Placemaker's response structure.

geographicScope and it contains the reference of the smallest place which best describes the document and its corresponding geographic coordinates.

4.2 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 preprocessing 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 [2], [4]. Due to hardware deficiency, later the C/C++ implementation also known as GibbsLDA++ [3] 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.

4.3 Graphic Generation

Given the outputs of the two tasks mentioned in Sections 4.1 and 4.2, 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.

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

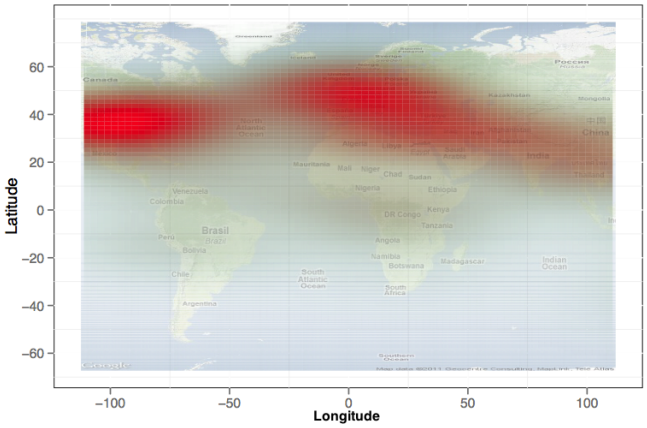

Figure 5: RCV1 Geographic document distribution.

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 [15]. To be noted that the smoothing of the data is achieved by using a Gaussian Kernel Estimator.

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.

5. EXPERIMENTAL EVALUATION

This section 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 section will describe the used dataset, the evaluation methodology, the conducted experiments, as well as discuss the achieved results.

5.1 Dataset

For the work's validation, it was conducted some experiments on the Reuters Corpus Volume 1 (RCV1), which is a collection of over 800,000 manually categorized newswire stories [11]. Each newswire story can be categorized over 55 topics, which are illustrated in Table 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 5.

5.2 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 3.3), and over space, using geographic maps displaying the geographic distribution of each topic generated with a Kernel Density Estimator (see Section 3.4).

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 2: RCV1 list of topics and respective document distributions.

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 6. Also, its geographical distribution covers the same geographic regions as the RCV1 dataset as illustrated in Figure 7. 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 3 illustrates some examples of

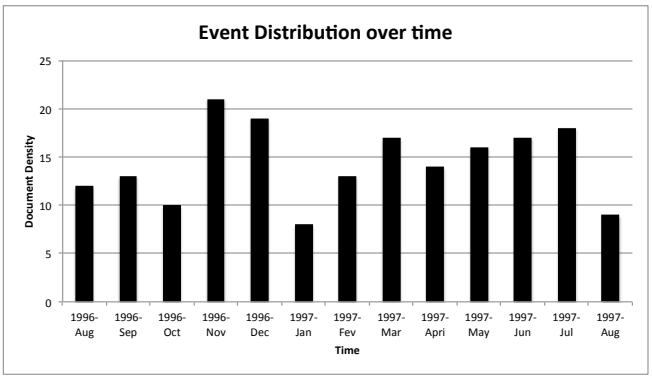

Figure 6: Temporal distribution of the set of important events.

this set of important events.
For the temporal topic analysis, we used the mentioned techniques in Section 3.3 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* 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 [9]. 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 3.4 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.

5.3 Experiments

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

5.3.1 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 2) and the automatically generated topics from the Latent Dirichlet Allocation algorithm (LDA topics).
The Latent Dirichlet Allocation classified the collection in

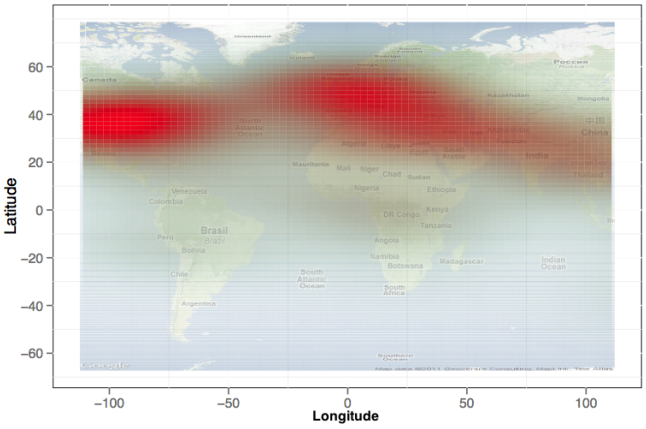

Figure 7: Geographic distribution of the set of important events.

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 **??**, 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. 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.

5.3.2 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 8, 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 9, proving once more the successful topic recreation of the Latent Dirichlet Allocation algorithm.

Another unique pattern that these temporal document distributions display are the *bursts* (see Section 3.3 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 10, the temporal document

Date	Event Description
20-Aug-1996	A thousands-large protest in Seoul calling for reunification with North Korea is broken up by riot police.
21-Aug-1996	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	The Big 12 Conference is inaugurated with a football game between Kansas State University and Texas Tech University in Manhattan, Kansas.
3-Sep-1996	The U.S. launches Operation Desert Strike against Iraq in reaction to the attack on Arbil.
14-Sep-1996	Alija Izetbegovifá is elected president of Bosnia and Herzegovina in the country's first election since the Bosnian War.
5-Feb-1997	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	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	In Lacey Township New Jersey high school senior Melissa Drexler kills her newborn baby in a toilet.
13-Jun-1997	A jury sentences Timothy McVeigh to death for his part in the 1995 Oklahoma City bombing.
8-Jul-1997	NATO invites the Czech Republic, Hungary and Poland to join the alliance in 1999.
10-Jul-1997	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	Miguel Ángel Blanco is kidnapped in Ermua, Spain and murdered by the ETA.
1-Aug-1997	Steve Jobs returns to Apple Computer, Inc at Macworld in Boston.
2-Aug-1997	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	Korean Air Flight 801 crash lands west of Guam International Airport, resulting in the deaths of 228 people.

Table 3: Set of Important Events.

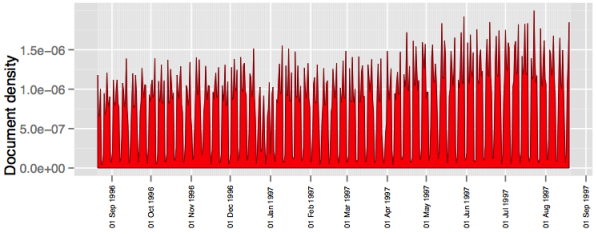

Figure 8: Document density over time for topic M11.

distribution of Topic *Disasters and Accidents* contains 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

Topic 7	Topic 10	Topic 24	Topic 25
health	military	stock	party
food	army	million	goverment
drug	goverment	share	election
care	forces	market	minister
medical	troops	closed	opposition
study	war	percent	prime
hospital	rebels	trade	elections
found	force	worth	leader
british	zaire	day	president
people	refugees	close	national
heart	fighting	company	parties
university	rebel	total	democratic
disease	people	high	support
caused	capital	hands	country
drugs	president	capital	general
treatment	peace	american	people
mother	soldiers	orders	leaders
human	mobutu	big	congress
said	aid	small	socialist
research	country	bought	presidential

Table 4: 20 most frequent words of some generated LDA Topic

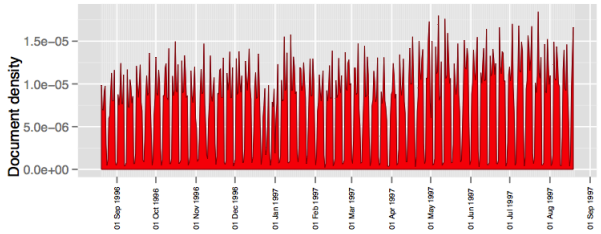

Figure 9: Document density over time for topic 24.

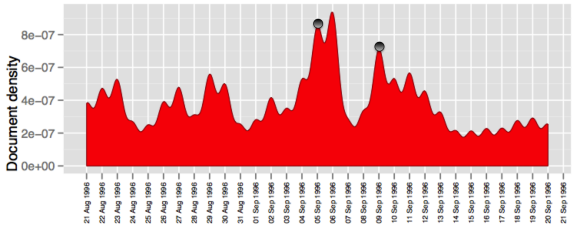

Figure 10: Document density over time for topic GDIS.

about an important event. For example this topic 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 illustrated in Figure 11, we can see that most of the important events of the topic *Elections* correspond

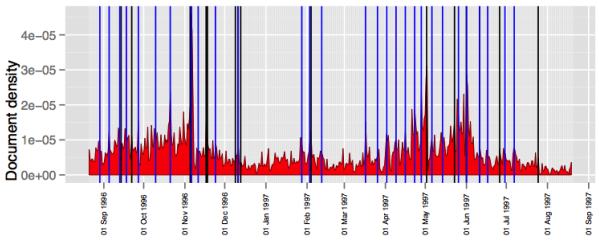

Figure 11: Document density over time for topic GVOTE.

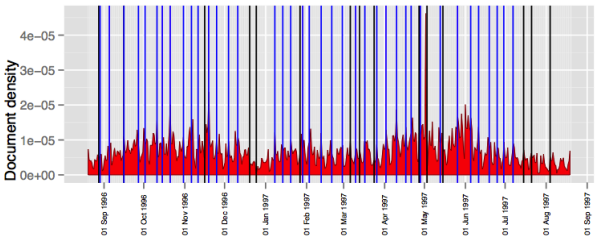

Figure 12: Document density over time for topic 25.

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.

5.4 Geographic Topic Analysis

The geographic topic analysis is a rather easier analysis to visualy evaluate. By using the previously described evaluation method in Section 5.2, 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 13 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 14, 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 the geographic document distribution of this LDA topic is less accurate to geographically pinpoint an important event then its corresponding manual assigned topic.

6. 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-

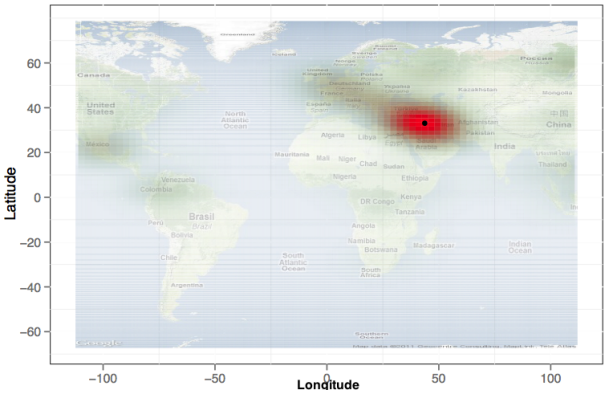

Figure 13: Document density over space for topic GVIO.

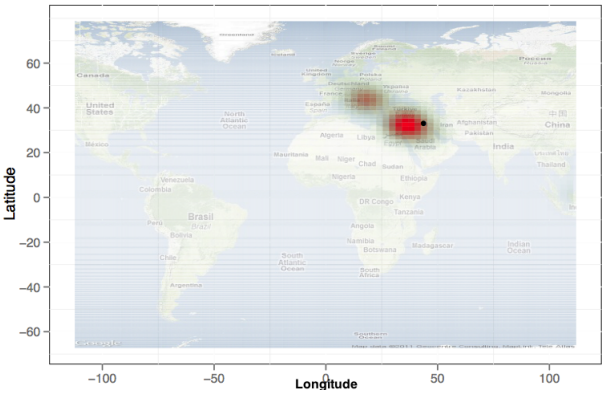

Figure 14: Document density over space for topic 1.

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.

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, such as:

- Integration of the Geofolk model in the topic analysis of newswire documents [12]. 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 [14]. This model could be used to understand the potential interrelation between topics over time and space.

7. REFERENCES

[1] J. Allan, editor. *Topic detection and tracking: event-based information organization*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 2003.

[3] H. Carlos, X. Shi, J. Sargent, S. Tanski, and E. Berke. Density estimation and adaptive bandwidths: A primer for public health practitioners. *International Journal of Health Geographics*, 2010.

[4] J. Chang. *Collapsed Gibbs sampling methods for topic models*, 2011.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 1977.

[6] J. Dobson, P. Coleman, R. Durfee, and B. Worley. Landscan: a global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing*, (7), 2000.

[7] W. Gilks and D. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC, 1996.

[8] M. Hurst. Temporal text mining. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, 2006.

[9] W. C. Lixin Gong and Y. A. Chen. *Protein Mass Spectra Processing*, 2011.

[10] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[11] T. Rose, M. Stevenson, and M. Whitehead. 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*, 2002.

[12] S. Sizov. 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, 2010.

[13] M. Steyvers and T. Griffiths. *Probabilistic Topic Models*. Lawrence Erlbaum Associates, 2007.

[14] M. Wahabzada, Z. Xu, and K. Kersting. Topic models conditioned on relations. In *Proceedings of the European Conference on Machine Learning*, 2010.

[15] H. Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.