Keyword and Event Extraction for Thematic Map Retrieval from Indonesian Online News Site

A Dewandaru, I Supriana and S Akbar
School of Electrical Engineering and Informatics
Institut Teknologi Bandung, Indonesia

Abstract. Online news sites provide great deal of information that may be extracted and mined to create Geographical Information Retrieval (GIR) system that provides thematic map based on user query. The task requires extracting generic events and its attributes from the news corpus. This event extraction requires the keywords and topics that can help increase the accuracy of the extraction and retrieval process. We prepare a large online news corpus and compared the Latent Dirichlet Allocation (LDA) and CBOW and skip-gram model to help providing base thematic keywords which can assist the extraction process on the intended GIR system. LDA is better in terms of semantic relatedness and the CBOW and skip-gram is useful for providing semantic similarity.

1. Introduction

The staggering growth of internet data is followed by the dominance of textual retrieval system to aid search. However, the extraction and retrieval of thematic maps from web corpus is an interesting task that is still not well explored [1]. The purpose of the task is providing infographic in form of thematic geographic map that is relevant to the user's query. In other words, it is an automated Geographical Information Retrieval (GIR) system that processes data from web and serves geographical content to users. Essentially it is a novel type of search engine, in which –instead of returning set of ranked documents– it returns a thematic map relevant to the query. The basic definition of many GIR follows the generic architecture and components illustrated in Figure 1.

On the diagram, the Extraction process (inside the IE Process box) deals with identifying and filtering information from the source web documents (such as named entity), notably handling the various formats and ambiguities, and also tackle the complexity of mostly less structured free text in form of natural languages. Even if we restrict the source to news sites (as the main focus of this paper) there are still a lot of open problems in the extraction process.

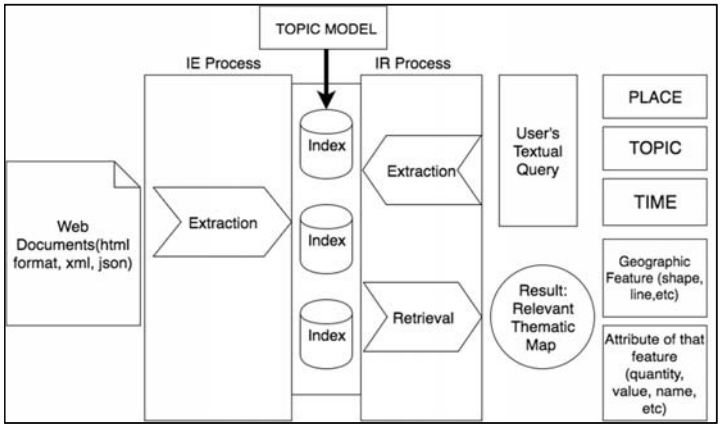

Figure 1. GIR System Architecture for Thematic Map Extraction & Retrieval from Web

Taking it to the greater level of difficulty, the task pretty similar to the machine reading or text understanding endeavour, exemplified by works [2,3]. However, we would like to underline the key scopes here that: 1) the text needs to have geographic reference(s) and 2) the result is a thematic map of a geographic scope that's relevant, which need not to provide direct answer from the user's query as in QA based system. The paper will explain the target output and how to deal with less-structured input in news documents. As later will be shown, we found that topic modeling will help in finding important keywords for extraction.

As stated in the earlier part, we would like to aim in creating GIR system that is able to output thematic map in response to user's query as in the example below (**Figure 2**). The map was published by Indonesian formal body (BNPB) and is projecting flood data in Jakarta in particular time range. The map was color-coded, which was further explained by a brief legend: blue area indicates a flooded area, while the light yellow showed clear area.

Such thematic maps were (typically) prepared manually using help of GIS by a formal or institutional body which has predefined editorial process and standards. We are aiming to create similar output automatically, tapping the techniques available in Machine Learning and Text Mining. However, instead of having a structured input such as database systems, we are interested in utilizing the world wide web as a huge and recent source of data and information. The inputs are taken from web documents instead of sampling from the field reports. Hence, the challenge is that it must be able to extract key information from the various types of web documents that is served in various formats.

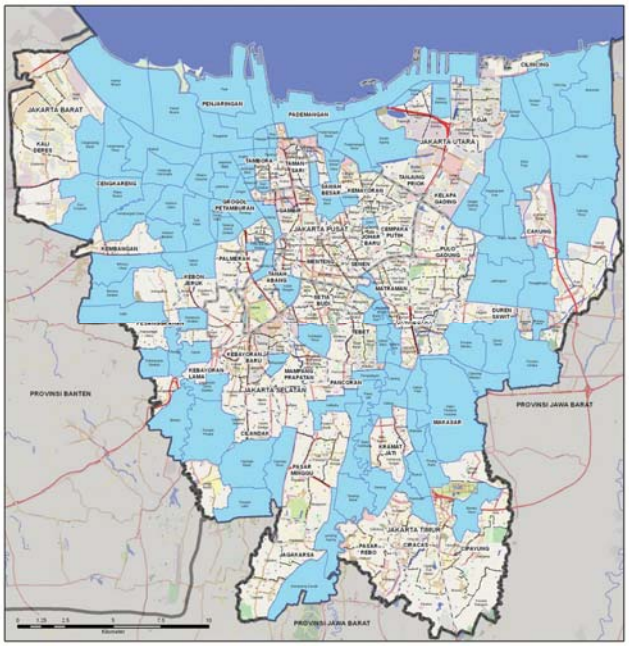

Figure 2. Thematic Map of Jakarta Flood 18th Jan 16

2. System inputs

There are myriad types of information and structure in the web documents that is suitable for thematic map extraction, i.e. such as formal body that publishes statistics regularly, but it may not match the timely update and coverage offered by social media and online news sites. However, we think that social media typically are not accurate enough for this purpose. Thus, in this paper we start with focusing on the (rather in the middle/balanced) news sites, especially in terms of the update frequency and trustworthiness due to a specific editorial process [4].

Table 1. Web data assumptions

Aspects	**Social Media**	**Formal Body**	**News Site**	**Wiki**
Update frequency	Fast	Slow	Medium	Slow
Information that can be mined	Events, preferences	Aggregate/ Instance statistics	Events	Events, Aggregate Statistics
Coverage	Broad	Narrow	Medium	Broad
Trustworthiness	Low	High	Medium	High
Format	Natural Language	Tabular HTML, DIVs	Natural Language	Natural Language, Tabular HTML
Example Sites	Facebook, Twitter	CIA World's Factbook	DailyMail, CNN	Wikipedia. org

Let us take a look on a sample news document (Figure 3) offered by the online news sites to get more details into its nature. Typically, news articles are about events and its attribute. By simply reading at the text, and by making use of the important keywords, it is easy for average adult human to

infer the 5W1H (what, where, when, who, why, how) attributes of an event presented in the article. However, it turns out to be a challenging task and still open problem for machine learning as we will shortly.

Dark sky, heavy rain and turbulent wind come across Jakarta. Sheds of water gushed out, as had happened in Kuningan area. Even when the water depth level reaches 30 cm, drivers were still trying to get across.

The report from detik.com, the rain started since 1:40 PM local time, Friday (29 February 2008) making a 30 cm watersheds in front of KPK building, situated at Jl HR Rasuna Said, Jakarta.

This had caused the traffic jam to direction of the Kuningan area. Drivers were slowing down when trying to get through. Meanwhile in the reverse side to Halimun, the flood caused vehicle queuing as far as 500 meters.

Figure 3. Translated Sample News Article (from http://news.detik.com/berita/902053/hujan deras-dan-angin-kencang-kuningan-banjir-30-cm)

3. Identifying keywords to extract events

An obvious problem that we face is that it is common to have multiple subtopics within a single news article. In the example given (Figure 3), the first portion (roughly first and second paragraph) was talking about the flood event. The second portion was about the traffic jam as the implication of the flood. A proper event extraction by a human annotator for that single article should give result similar to the following table:

Table 2. Extracted Values form Single Document Sample

Topics #	Variable	Value
1	Event	Flood
	Time	29/2/2008
	Place	Jakarta, Kuningan, Jl. HR Rasuna Said
	Depth	30 cm
2	Event	Traffic Jam
	Time	29/2/2008
	Place	Jakarta, Kuningan, Halimun
	Queue Length	500 m

Note that the example presented is only a simple case of a news event and with a specific domain (i.e. disaster news). With a specific domain, an information extraction wrapper or supervised wrapper induction system is easier to craft manually or trained, notably by identifying keywords term which plays some semantic role in the typical sentence. For example, Tanev et. al [5] are using set of keywords (kill, shoot) to help construct handcrafted rules which will handle with specific domain of disaster and violence.

Similarly, [6] had prepared manually annotated corpus in Bahasa Indonesia to train some supervised machine learning classifiers (Adaboost.M1 and C4.5) to perform event extraction in output form of 5W1H. The learning is based on a limited number of only 90 annotated news articles. We think that it is not possible to cover a huge repertoire of articles in the real data without training pretty large portion of a manually annotated samples (the training would usually assign the statistical significance of some keywords).

In the work of [7], the identification of key terms are done automatically using TF-IDF metric that is monitored daily with some simple rules: 1) it has to be present in more than two documents and 2) it is ranked via additional “novelty” metric that takes care of the currency of the term (we compare this

in modern parlance as "trending topic" concept). However, the method only extracted hot terms as the main topic from the news article; whereas in our case we need to get down to the attributes keyword (such as queue, water depth, etc.) and its relevant value (500 m, 30 cm).

In a real world scenario –e.g. in the context of a typical news sites– we are facing multi-faceted, intertwined and yet multi-topics (hundreds) each with its distinct sets of keywords that arise from multiple domains of news articles to extract from, and the number of tokens exposed is pretty huge (in hundred-thousands tokens, as will be confirmed later on based on our crawled data). Hence it is not a simple task, and requires great deal of manual work if done solely using supervised algorithms.

Recall that the GIR described in the introduction needs to have a great emphasis in the ability of returning relevant maps pertaining to user's query. Other GIR had specified different query formats. As an example, the SPIRIT project [8] for example had offered the query format as triplet of theme (e.g. schools), spatial relationship (such as near, within, etc.) and location string. Note the theme that is being queried as important keyword in the search query there.

In our designated system we wanted to be able to extract such thematic information, and even provide ability to extract intangible object (not just some building or premises, but such as events as in floods, earthquake, or spatially related info such as GDP statistics) and enable users to query these information needs (e.g. floods in Jakarta 2014, or World's GDP 2015). The triplet of query format would form as <theme><location><time> in an unspecified sequence. The examples of queries are given in **Table 3**.

Table 3. Example of queries and expected output

Query example	Output
"Flood Jakarta 2014"	Summarized map of flood in Jakarta within 2014
"World GDP 2015"	Map of world countries' growth domestic products statistics
"Earthquake Asia 2010"	Map of earthquake centers that happens in 2010 in Asia Pacific.

To tackle this problem, we suggest the usage of unsupervised algorithms to assists us in learning or extracting important topics keywords that would eventually useful in extracting proper values based on the theme or topics. Probabilistic topic modeling framework falls exactly as a good option for this.

4. Topic modeling for keyword extraction

Topic modeling is a task of capturing hidden relations between terms and finding clusters terms in the corpus. It can be traced from a very simple TFIDF metric, which can be used to gauge term significance by accounting on the term frequency and the term specificity. Deerwester et. al [9] developed Latent Semantic Indexing (LSI), which tries to capture hidden pattern not directly visible in the co-occurrence of words using Singular Value Decomposition (SVD) applied to term-document matrix. Later this work was developed further and extended by basing it under probabilistic framework called Probabilistic Latent Semantic Indexing (PLSI) by Hoffman [10]. This provided an aspect model and was a significant upgrade over the former.

The seminal Latent Dirichlet Allocation (LDA) by Blei et.al. [11] supposed to fix the overfitting problem and the need to linearly grow the parameter as the number of documents increase in PLSI. Unlike some traditional IR clustering approach (e.g. agglomerative clustering or k-means), Both LDA and PLSI does a soft-clustering in which they do not form an absolute partition. These architectures offer a mixture model whereby a words formed from a mixture of term components (we can imagine a cluster of words) that represents topic. It models the possibility that a document consisted of several topics at the same time.

In a mixture model, an element can be a member of more than one cluster with varying degree of membership [12]. In the case of LDA and PLSI, a topic is nothing more than virtual "set" of terms, of which each term is having degree of membership to any topic to all other topics. Thus a term belongs to more than one topic, forming many to many relationships.

As an example, we can model topics of two different events (flood and quake) as a mixture components of terms, each having an attached probability value indicating how likely it forms that topic. Note that "victim" term are both highly probable member of Flood and Quake topic. Any terms in the processed corpus are actually having a probability to any topic, but for practical purpose some cut-off values are being used to prevent low probability terms to appear.

We believed that if these clusters of important keywords discovered accurately, it can be a very useful information that can assists many tasks such as classification, extraction and retrieval. We will show later on the results found are pretty consistent with our expectation.

5. Latent Dirichlet Allocation

The LDA model is a widely topic modeling framework and is based on a sound probabilistic graphical model framework that itself belongs to a family of Bayesian Network. Its use on GIR literature is still minimal however, and we are interested in using it due to its ability to extract keywords and topics at once.

In this model, the probability assumptions is encoded in form of directed acyclic graph[13], where each node represents a random variable which can be observable or hidden. A hidden or latent node is not observable directly but can often be estimated from the visible ones. The directed edge of this graph represents dependencies between random variables (parent nodes has directed edge to its child nodes, the child is conditionally dependent on its parents).

To illustrate better, Figure 4 shows LDA model which is presented in a plate notation (or often called hierarchical model). Plate (drawn as rectangular boxes) notation is a visual extension which provides easy way to visualize duplication of nodes semantics. N and M in the graph are the quantifier variable, indicating how many nodes are duplicated inside the plate box. Thus, Z and W will be duplicated a number of N times.

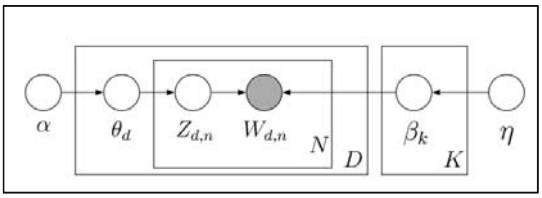

Figure 4. LDA graphical model with plate notation

From the graphic, we can also say that the nodes α, β, θ, Z, W are random variables that is related one another via the edge annotation. For example, W is conditioned (depends) on Z and so forth. W is the only random variables that is observable (shaded). In the LDA case it is the word inside documents formed from N words.

As with many Bayesian inference models, LDA is a generative models, in which it tries to construct a joint probability distribution[14] that models the entire corpus and able to "generate" new documents. From the model standpoint it is simply multiplication result of each instantiated node given its parent's value. It can be used to create a new document that is having similar distribution as the learned ones. Note that LDA uses bag-of-words unigram assumptions, thus it does not care about the sequence of words as it uses bag of words representations.

$$P_B(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P_B(X_i|\pi_i)$$

Figure 5. Joint Probability Distribution of A General Bayesian Network

Using the formula from the generative model, and by observing the training documents, the parameters to the network will be approximated or estimated. One problem is that it is often the case that the formula for the equation requires calculating integral that is intractable to compute and needs to be simplified further. Parameter can be used by some methods. Researchers use Gibbs Sampling (based on stochastic simulation) or some Expectation Maximization (EM) algorithm (including Variational). Both EM and Variational are deterministic and are not based on stochastic simulations. These learned parameters will determine topic structure: the terms and its memberships within the topics.

A final note on these method is that it requires a great deal of data to train with. They are unsupervised however, so minimum human participation is anticipated. So we are going to proceed by describing the acquisition of the training data from two of the biggest Indonesian news sites, which should produce a good enough corpus for these model.

6. Skip Gram and Continuous Bag of Words Models (Word2Vec)

The recent phenomena of massive generation of text via the internet motivates models and algorithms to be scalable, faster and better in handling data, at even more than hundreds of million of tokens. Also, the performance improvements of machine learning inference techniques enable more complex model to be trained within reasonable time.

The language model used within LDA context is the unigram known as bag of words. This model discards the information sequence that may arise within document. It is simpler and faster to train and handle, but the prediction performance or the quality of the model may suffer if compared to more complex model.

Mikolov et.al [15] introduced a novel approach which seen by many as promising direction in language modelling. The skip-gram model proposed is an extension of n-gram, which, instead of modelling a contiguous series of tokens in the model, it allows one or more “skips” between the token involved. It is also introduced a novel way to measure the “cleverness” of the algorithm in noticing patterns such as answering question like this: “Jakarta is to Indonesia is like Berlin to Germany”.

7. Experiment

We are going to describe main steps in this experiment. The manual annotation of the corpus is not yet used but article tagging and application has been used as proof of concept. The main objective is to check whether we can use keywords learned from the unsupervised learning to serve in topic extraction and retrieval. The following methods are done:

1. Obtain good amount of news corpus in Bahasa Indonesia.
2. Unsupervised training using LDA and Word2Vec algorithm from the entire set of documents within the corpus.
3. Analysis the keyword against sample documents.
4. Analysis the similarity of sample keywords.

#554:
banjir, air, sungai, warga, hujan, cm, tergenang, rumah, surut, deras, jalan, ketinggian, wilayah, pompa, meter, terendam, kawasan, menggenangi, lokasi, meluap, airnya, tanggul, aliran, mengguyur, titik, mencapai, daerah, akibat, luapan

#517:
kendaraan, tol, kemacetan, arah, jalan, macet, lintas, arus, jalur, lancar, pengendara, melintas, padat ,jam, roda, kilometer, mobil, petugas, antrean, pintu, masuk, titik, kawasan, motor ,bergerak, lokasi, alternatif, ditutup, pagi

#209:
gempa, bumi, mengguncang, tsunami, berkekuatan, kerusakan, kilometer, akibat, warga, barat ,korban, petang, wilayah, rumah, peringatan, pantai, pusat, bencana, orang, jiwa, dampak, laporan, tewas, stroke, sebelah, berolahraga, dilaporkan, dilansir, tersumbat, terserang

Figure 6. Snippet of LDA topic model learned, topic #554 (flood) and #517 (traffic jam) and #209 (earthquake)

The first target of news corpus was not directly available, so it was necessary to create such. We crawled news articles from www.detik.com and www.kompas.com, both were the leading news portal in Indonesia at the moment. We assume they hosted a great number of articles and having a relatively good editorial quality. Both sites are having archive index page which sort articles based on date of publication.

The crawling was done within a two level scopes: the index and the detail article. Scraping had been done for the detail article page based on CSS selectors (manually crafted wrappers) which has been adapted for both sites.

The crawling process was directed to gather title, content, and the URLs of the articles from January 2013 to May 2016. It resulted 271.979 articles in Bahasa Indonesia (Indonesian) from www.detik.com and 330.249 articles from www.kompas.com. The result of this crawling was cleaned from HTML tags and irrelevant scripts. We did not apply stemming as we would like to learn as much word forms as possible from the corpus. A stop words filtering was applied based on simple dictionary in Bahasa Indonesia. We also applied standard TF-IDF transformation to downplay the less specific but often appearing terms. Total unique terms count extracted for www.detik.com reaches 168,865 tokens.

Afterwards, the LDA model was constructed and the unsupervised learning method to extract the topics were executed based on the cleaned corpus. The training is done using online LDA algorithm by [16] via gensim python library[17]. We were playing with several combinations of number of topics (k = 100, 200, 500, 600) and number of terms per topic (t = 10, 20, 30) before settling to k = 600 and t = 30 which we think best capturing the size of the data.

The Word2Vec family model was trained using both skip-gram and continuous bag of words (CBOW) model using line-by-line training sequence. The maximum distance between current and predicted word is set to 5 and the dimensionality of the feature vector is 400. The Word2Vec training process took some minutes and is much faster compared to the LDA.

8. Results and Discussion

8.1 LDA

We had trained a model of LDA with the obtained corpus in an unsupervised manner. We manually picked three notable clusters (topics) out of 500 described on Figure 6 to serve as example and applied for one article. With that configuration, the LDA seems to be able to cluster semantically related terms pretty well and suited to our task of keyword extraction. When this result was applied on real articles example cited above (Figure 7 was the real article whose translation is presented on Figure 3), we found that many important keywords to be extracted from page already described by the clusters' term keyword. This serve as proof-of-concept of the utility of using topic modeling approach towards thematic map extraction GIR system.

The experiment showed some weaknesses however, in that we need to arbitrarily picked number of topics (repeated trials). Bayesian nonparametric techniques such as HDP [18] are able to infer the number of clusters automatically. Also, LDA is agnostic about geospatial data and temporal data. However some specialization on LDA in this direction has been made, e.g. [19][20] with Spatial LDA and 3S-LDA to incorporate geospatial and temporal data. LDA also does not regard sequential pattern nor works in the sentence level. It only views document as bag of words, not accounting sequence as in n-gram models. It helps in simplifying the model assumption but we think that it is very crucial in order to do proper extraction to reveal hidden patterns in sentence or sequence level as in Word2Vec.

Langit kelam, *hujan* deras dan angin kencang melanda Jakarta. Genangan *air* langsung bermunculan, seperti halnya di *kawasan* Kuningan. Meski *ketinggian air* mencapai **30** *cm*, pengendara tetap mencoba melintasinya.

Pantauan detikcom, *hujan* yang turun sejak pukul 13.40 WIB, Jumat (29/2/2008), membuat genangan air **30** *cm* di **depan kantor KPK, Jl HR Rasuna Said, Jakarta**.

Akibatnya *arus* lalu *lintas* mengarah ke Kuningan tersendat. *Pengendara* memperlambat laju *kendaraan* saat hendak melintasi genangan *air*. Sedangkan *arah* sebaliknya ke Halimun, *banjir* menyebabkan *antrean kendaraan* mencapai **500** *meter*.

Sementara *pengendara roda* dua memilih berteduh di halte-halte bus. Pengojek payung berwarna-warni pun laris manis.

Figure 7. Projection of the Extracted Keywords on A News Article from Topic #554 (flood) and Topic #517 (traffic jam). Bold terms indicate the values that should be extracted with regard to the event. Italicized keywords are related to flood and underlined words are related with traffic jam topic.

8.2 Word2Vec

The skip-gram and CBOW model offered interesting word similarity result. The example of this can be seen on **Table 4**. We noticed that LDA results are focused towards semantic relatedness while Word2Vec are more on semantic similarity. The Word2Vec result cannot be immediately used as keywords for extraction, however we are interested in its ability to produce synsets, hyponyms, and hypernims.

Table 4. Similarity of "Flood" (Ind. “Banjir”) (same word forms are omitted)

Word	Score
Cileuncang	0.668
Bandang	0.651
Kekeringan	0.611
Terendam	0.609
Rob	0.596
Longsor	0.584
Genangan	0.576
Luapan	0.559
Hujan	0.520
Plengsengan	0.503
Bencana	0.494

9. Conclusions and Future Works

Topic modeling framework with its unsupervised algorithms can be used to assist Event extraction process by providing priors over some important keywords. We plan to employ the set of keywords learned to assists the in-page information extraction process by commencing a web browser extension module development. The planned algorithm for extraction falls under the family of discriminative type (e.g. sequential Conditional Random Field), which will be able to make use of neighbourhood term observation data. For the topic and keyword extraction we plan to adapt the Bayesian approach of LDA, combined with the skip gram approach to fit better with online news domain in terms of tackling more on geospatial and temporal data.

REFERENCES

[1] Dewandaru A, Supriana I and Akbar S 2015 Evaluation on Geospatial Information Extraction and Retrieval : Mining Thematic Maps from Web Source *Information and Communication Technology (ICoICT), 2015 3rd International Conference on* (IEEE) pp 283–8
[2] Etzioni O, Banko M, Soderland S and Weld D S 2008 Open information extraction from the web *Commun. ACM* 51 68
[3] Etzioni O, Popescu A, Weld D S, Downey D and Yates A 2004 Web-Scale Information Extraction in KnowItAll (Preliminary Results) *Proceedings of the 13th international conference on World Wide Web Pages* pp 100–10
[4] Haristya S, Suwana F and Kurniana I 2012 THE CREDIBILITY OF NEWS PORTAL IN INDONESIA : *J. Commun. Stud.* 5 1–17
[5] Tanev H, Piskorski J and Atkinson M 2008 Real-time news event extraction for global crisis monitoring *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol 5039 LNCS pp 207–18
[6] Khodra M L 2015 Event Extraction on Indonesian News Article Using Multiclass Categorization 1–5
[7] Liu M, Liu Y, Xiang L, Chen X and Yang Q 2008 Extracting key entities and significant events from online daily news *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 5326 LNCS 201–9
[8] Purves R S, Clough P and Jones C B 2007 The Design and Implementation of SPIRIT : a Spatially- Aware Search Engine for Information Retrieval on the Internet *Int. J. Geogr. Inf. Sci.* 717–45
[9] Deerwester S, Dumais S T, Furnas G W, Landauer T K and Harshman R 1990 Indexing by latent semantic analysis *J. Am. Soc. Inf. Sci.* 41 391–407
[10] Hofmann T 1999 Probabilistic latent semantic indexing *Acm Sigir* 50–7
[11] Blei D M, Ng A Y and Jordan M I 2003 Latent Dirichlet Allocation *J. Mach. Learn. Res.* 3 993–1022
[12] Song M, Wu Y B and Klinger K 2009 *Handbook of Research on Text and Web Mining Technologies* vol II (IGI Global)
[13] Koller D and Friedman N 2009 *Probabilistic Graphical Models: Principles and Techniques*
[14] Faltin F and Kenett R 2007 Bayesian Networks *Encycl. Stat. Qual. Reliab.* 1 4
[15] Mikolov T, Corrado G, Chen K and Dean J 2013 Efficient Estimation of Word Representations in Vector Space *Proc. Int. Conf. Learn. Represent. (ICLR 2013)* 1–12
[16] Hoffman M, Blei D and Bach F 2010 Online learning for latent dirichlet allocation *Nips* 1–9
[17] Řehůřek R 2011 Scalability of Semantic Analysis in Natural Language Processing 147
[18] Teh Y W, Jordan M, Beal M and Blei D 2006 Hierarchical Dirichlet Processes *J. Am. Stat. Assoc.* 101 1556–81
[19] Wang X and Grimson E 2008 Spatial Latent Dirichlet Allocation *Adv. Neural Inf. Process. Syst.* 1–8
[20] Pan C-C and Mitra P 2011 Event detection with spatial latent Dirichlet allocation *Proceeding 11th Annu. Int. ACM/IEEE Jt. Conf. Digit. Libr. - JCDL '11* 349