HAL
open science

Wikipedia-based Semantic Approach for Tweet Contextualization

Meriem Amina Zingla, Chiraz Latiri, Yahya Slimani, Catherine Berrut

To cite this version:

Meriem Amina Zingla, Chiraz Latiri, Yahya Slimani, Catherine Berrut. Wikipedia-based Semantic Approach for Tweet Contextualization. KDDA - Knowledge Discovery ans Data Analysis, Nov 2015, Alger, Algeria. hal-01576443

HAL Id: hal-01576443

https://hal.science/hal-01576443v1

Submitted on 23 Aug 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Wikipedia-based Semantic Approach for Tweet Contextualization

Meriem Amina Zingla[1,2], Chiraz Latiri[3], Yahya Slimani[1], and Catherine Berrut[4]

[1] INSAT, LISI Research Laboratory, University of Carthage, Tunis, Tunisia
[2] Faculty of Sciences of Tunis, University of Tunis El Manar, Tunis, Tunisia
[3] University of Tunis El Manar, Faculty of Sciences of Tunis, LIPAH research Laboratory, Tunis ,Tunisia
[4] Grenoble Alpes University, LIG laboratory, MRIM group, Grenoble, France

Abstract. The tweet contextualization task aims at providing an automatic readable summary explaining a given tweet. As tweets are very short documents, bound to 140 characters, and not always written maintaining proper spellings, there is indeed a need for such a task. This article describes a semantic tweet expansion approach for the tweet contextualization task based on Wikipedia as an external knowledge source. This approach consists in two major phases, namely: the first is the generation of the candidate terms, from Wikipedia. While the second is the selection of the most-related terms. To achieve this latter, we propose a semantic relatedness measure based on the Explicit Semantic Analysis and association rules mining. The effectiveness of our approach is proved through an experimental study conducted on the INEX 2014 collection. Our results have outperformed of the runs issued from INEX 2014.

Keywords: Information Retrieval, Tweet Contextualization, Query Expansion, Explicit Semantic Analysis, Association Rules Mining.

1 Introduction

Twitter is a communication medium and a collaboration system that allows broadcasting short messages called tweets. In contrast to traditional blogs, the textual part of tweets is limited to 140 characters, submitted in real-time to report an idea, an actual interest, or an opinion [4]. The limit on the length of a tweet causes the well-known vocabulary mismatch problem, rendering the tweet hard to understand. Hence, to make it understandable by readers, it is necessary to find out its contexts.

The aim of the INEX (Initiative for the Evaluation of XML retrieval) tweet contextualization task is to provide, automatically, from a recent dump of Wikipedia, a context that explains a given tweet, to help the reader understand this latter. To achieve this, two systems are combined, Information Retrieval System (IRS)

based on the Indri[1] search engine and Automatic Summarization System (ASS) based on an efficient summarization algorithm created by TermWatch[2]. While the IRS extracts, from the Wikipedia document collection, a set of relevant documents for a given tweet, the ASS selects the most relevant passages from the extracted documents. These passages, not exceeding 500 words, define the context of the tweet. A baseline system composed of an IRS and an ASS has been made available online[3].

In this paper, we define the tweet contextualization task as a query expansion issue. We are only interested in the textual part of the tweets, hence, we consider these latter as queries. The aim is to enhance the quality of a tweet for the baseline system, since it has a direct impact on the context quality. To accomplish this, we propose a semantic approach that consists of two phases, the first one is based on Wikipedia as an external knowledge source, and allows to generate candidate terms for a given tweet. The second one is based on Explicit Semantic Analysis [8] and association rules mining [1], and calculates the semantic relatedness score between the tweet and the candidate terms. The relatedness score is used to rank the candidate terms in order to select the best ones for the tweet. We opted to use Wikipedia because it is currently the largest knowledge repository on the Web [9]. It is available in dozens of languages, with its English version being the largest increased every day with over 800 new articles. We claim that the use of this huge source is fruitful for the tweet contextualization task allowing a massive knowledge representation of a tweet.

Our proposed approach for tweet contextualization offers an interesting solution to obtain relevant context . This mainly relies on an accurate choice of the added terms to an initial tweet. Interestingly enough, tweet contextalization takes advantage of large text volumes provided by wikipedia articles by extracting semantic information.

The remainder of this paper is organized as follows: In section 2, our work is put in the context of related works while section 3 introduces the formal explanation of our problem. Section 4 gives a detailed description of our tweet contextualization approach and section 5 presents our experimentations and results. Finally, section 6 is dedicated to the conclusion of this work and gives future works.

2 Related Works

2.1 Tweet Contextualization

Despite the fact that the idea to contextualize tweets is quite recent, there are several works in this context. Recently, authors of [7] proposed a method based on the local Wikipedia dump, they used the Term Frequency-Inverse Document Frequency TF-IDF cosine similarity measure enriched by smoothing from local context, named entity recognition and Part-Of-Speech weighting presented at

[1] http://www.lemurproject.org/indri.php
[2] http ://data.termwatch.es
[3] http://qa.termwatch.es/data

INEX 2011. They modified this method by adding bigram similarity, anaphora resolution, hashtag processing and sentence reordering. The sentence ordering task was modeled as a sequential ordering problem, where vertices corresponded to sentences and sentence time stamps represented sequential constraints. They proposed a greedy algorithm to solve the sequential ordering problem based on chronological constraints.
While in [5], authors used a method that allows to automatically contextualize tweets by using information coming from Wikipedia. They treated the problem of tweets contextualization as an automatic summarization task, where the text to resume is composed of Wikipedia articles that discuss the various pieces of information appearing in a tweet, one of the limitations of this approach is that the number of Wikipedia articles used to extract the candidate sentences is set manually. They explore the influence of various tweet-related articles retrieval methods as well as several features for sentence extraction. Whereas, in [6], authors added a hashtag performance prediction component to the Wikipedia retrieval step. They used all available tweet features including web links which were not allowed by INEX's organisers.
In [13], authors used an automatic summarizer named REG. based on a greedy optimization algorithm to weigh the sentences. The summary is obtained by concatenating the relevant sentences weighed in the optimization step. In [15], authors used Latent Dirichlet Analysis (LDA) to obtain a representation of the tweet in a thematic space. This representation allows the finding of a set of latent topics covered by the tweet, this approach gives good results for the tweet contextualization task. Authors in [21] use the association rules between terms to extend the tweet, they project the terms of tweet on the rules' premises and add the conclusions to the original tweets. Finally, in [19] authors developed three statistical summarizer systems the first one called Cortex summarizer, that uses several sentence selection metrics and an optimal decision module to score sentences from a document source, the second one called Artex summarizer, that uses a simple inner product among the topic-vector and the pseudo-word vector and the third one called Reg summarizer which is a performant graph-based summarizer.

2.2 Query Expansion

Several works in the literature are proposed for the query expansion task, such as in [17] where the authors proposed a novel semantic query expansion technique that combines association rules with ontologies and Natural Language Processing techniques. This technique uses the explicit semantics as well as other linguistic properties of unstructured text corpus, it incorporates contextual properties of important terms discovered by association rules, and ontology entries are added to the query by disambiguating word senses. In [16], authors proposed to perform an initial retrieval of resources according to the original keyword query, the proposed process is divided into three main steps. In the first step, all words closely related to the original keyword are extracted based on two types of features: linguistic and semantic. In the second step, the introduced linguistic and

semantic features are weighted using learning approaches. In the third step, they assign a relevance score to the set of the related words. Using this score we prune the related word set to achieve a balance between precision and recall. In [11], authors addressed query expansion by considering the term-document relation as fuzzy binary relations. Their approach to extract fuzzy association rules is based on the closure of an extended fuzzy Galois connection, using different semantics of term membership degrees. Authors in [18], propose a semantic approach that expands short queries by semantically related terms extracted from Wikipedia, they incorporate the expansion terms into the original query and adapt language models to evaluate the expanded queries.
Query expansion techniques are also used for microblog retrieval, authors in [2], for example, used external corpora as a source for query expansion terms. Specifically, they used the Google Search API (GSA) to retrieve pages from the Web, and expanded the queries employing their titles. In [12], authors proposed a twitter retrieval framework that focuses on topical features, combined with query expansion using pseudo-relevance feedback (PRF) to improve microblogs retrieval results.

In this work, we propose to use a query expansion technique for the tweet contextualization task. Similar works have already proposed the use of the expansion technique for the target task ([15] and [21]), but they suffer from a major drawback. This drawback comes from the fact that these approaches did not include a term ranking phase. The absence of this ranking step resulted in noisy queries meaning that an extended query contained terms that did not relate to the original one. Taking these weaknesses into account, we propose to enhance our query expansion technique so that it will be composed of two phases, namely: the candidate terms generation, from Wikipedia, and the selection of the best ones according to their relatedness score. This selection step ensures that we will extend the query with related terms only. To calculate the semantic relatedness score we propose a new measure basing on ESA and association rules mining.

3 Basic Notions and Preliminaries

After introducing some notations, we state the formal definitions of the concepts used in the remainder of the paper. In this respect, we shall use in text mining field, the theoretical framework of Formal Concept Analysis (FCA) presented in [10]. First, we formalize the tweet contextualization task. So, the tweet contextualization task is concerned with contextualizing a set of n tweets $\Gamma = \{tw_1, \ldots, tw_n\}$ using a collection of m Wikipedia articles $\Sigma = \{d_1, \ldots, d_m\}$ by providing a context c_i for each tweet $tw_i \in \Gamma$. For a given tweet tw_i we retrieve a sub-set Σ_p from Σ of the relevant articles, then we select the most relevant passages from the Σ_p. These passages define the context c_i.

3.1 Tweet Representation

We consider tweets as a short documents. We represent a tweet as bag of words, we do not use it directly in our proposed approach, but apply a preprocessing

step first, wich removes all stop-words and twitter's specific stop-words such as (RT, @username, #). Formally, we have:

$$tw_i = \{wt_j\} \tag{1}$$

where
wt_j : *is a word in the tweet* tw_i;
$i, j \in \mathbb{N}$.

3.2 Tweet Context Representation

For a given tweet tw_i, a context c_i is a concatenation of passages from the articles in Σ_p sub-set. a passages is composed of set of words. Formally, we have:

$$c_i = \{wc_1 \dots wc_k\} \backslash k \leq 500 \tag{2}$$

where
wc_k : *is a word in* c_i associated to the tweet tw_i;
$k \in \mathbb{N}$.

3.3 Association Rules Mining

An association rule between terms R is an implication of the form R: $T_1 \Rightarrow T_2$, where T_1 and T_2 are subsets of $\mathcal{I}$, where $\mathcal{I} := t_1, ..., t_l$ is a finite set of l distinct terms in the Wikipedia document collection and $T_1 \cap T_2 = \emptyset$. The termsets T_1 and T_2 are, respectively, called the *premise* and the *conclusion* of R. The rule R is said to be based on the termset T equal to $T_1 \cup T_2$. The *support* of a rule R: $T_1 \Rightarrow T_2$ is then defined such that:

$$Supp_{R(T_1,T_2)} = Supp(T), \tag{3}$$

while its *confidence* is computed such that:

$$Conf_{R(T_1,T_2)} = \frac{Supp(T)}{Supp(T_1)}. \tag{4}$$

An association R is said to be *valid* if its confidence value, *i.e.*, $Conf_{R(T_1,T_2)}$, is greater than or equal to a user-defined threshold denoted *minconf*. This confidence threshold is used to exclude non valid rules.

4 The Proposed Approach for Tweets Contextualization

The tweet contextualization task is used to extract the context of a given tweet. The main goal is to enhance the quality of this context, *i.e.*, ensuring that the context contains adequate correlating information with the tweets and avoiding the inclusion of non-similar information. To reach this goal, we propose a semantic approach based on Wikipedia as a Semantic Source (*cf.* Figure 1) that is divided into two major phases, the first one, namely the Tweet Expansion, consists in generating candidate terms for a given tweet. The second one, namely Term Ranking, consists in selecting the most related terms and adding them to the tweet.

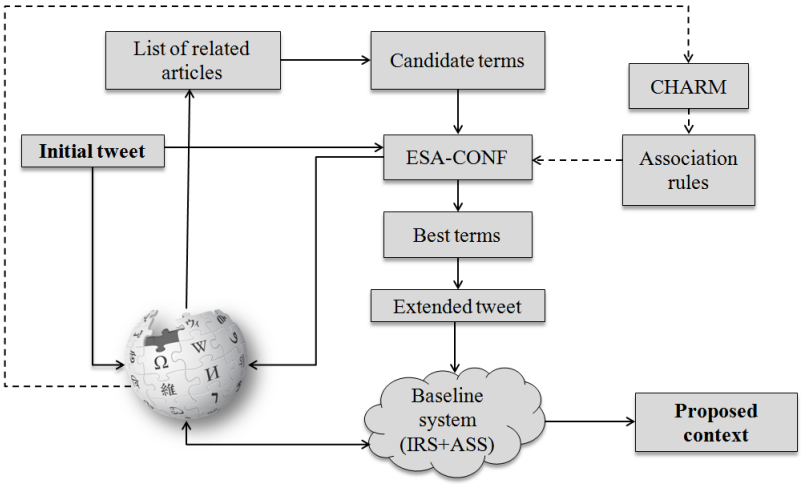

Fig. 1. The proposed approach architecture

Tweet Expansion: It consists in exploring the Wikipedia articles related to the tweet. The nouns appearing in these articles' first sentences represent the set of candidate terms of the tweet. We note that an article's first sentences are called *definitions*. To achieve this, we use some heuristics:

- Given a tweet tw_i, first, we search for all articles that correspond to each word $wt_j \in tw_i$ in Wikipedia, using WikipediaMiner[4][14] for example, for Volvo we find the following articles: Volvo, Volvo Cars, Volvo Buses, Volvo Trucks, etc.
- We select the most referenced articles, these latter follow a predictable layout, which allows to provide short definitions of wt_j by extracting the corresponding article' definition. For the previous example, we retrieve the following definition :
  "The Volvo Group is a Swedish multinational manufacturing company headquartered in Gothenburg." from the Volvo article that has the highest relatedness with the query.
- We annotate these definitions using TreeTagger which is a tool for annotating text with part-of-speech and lemma information. The choice of TreeTagger was based on the ability of this tool to recognize the nature (morpho-syntactic category) of a word in its context. TreeTagger uses the recursive construction of decision trees with a probability calculation to estimate the part of speech of a word.
- We extract the nouns from these annotated definitions. These nouns form the set of candidate terms among which the terms that will be added to the tweet are selected.

[4] http://wikipedia-miner.cms.waikato.ac.nz/

Term Ranking: This phase consists in ranking the candidate terms according to their relatedness to the given tweet, and selecting the best ones to add to the initial tweet. To achieve this, we propose a new semantic relatedness measure (ESA-CONF) that combine the Wikipedia-based Explicit Semantic Analysis (ESA) measure and the association rules' confidence value.
The following steps detail the *ESA-CONF* measure:

- Generating the association rules between terms from the Wikipedia articles, using an efficient algorithm, namely: CHARM[5][20], and use their confidences

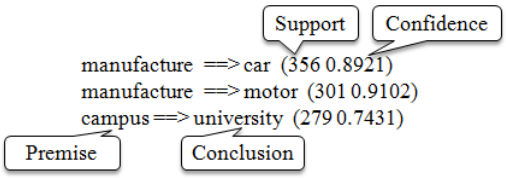

Fig. 2. An Example of Association Rules Extracted by CHARM

(*cf.*Figure 2) to calculate the new score of relatedness such that:

$$ESA - CONF_{tw_i,w} = \begin{cases} \alpha\, ESA_{tw_i,w} \;+\; (1-\alpha)\, Conf_{R(wt,w)} \;\; if\, \exists\, R(wt,w); \\ ESA_{tw_i,w},\; otherwise. \end{cases} \tag{5}$$

Where

- $ESA_{tw_i,w}$ is the score of relatedness between the tweet tw_i and the candidate term w calculated by *ESA*, we used the *ESA* implementation described in [8].
- $Conf_{R(wt,w)}$ is the confidence of the rule R that express the association between the candidate term w and a word in tweet wt.
- α is a weighting parameter $\in [0,1]$.

- calculating the *ESA-CONF* relatedness score between the tweet and each of the candidate terms.
- selecting the most related terms according to their *ESA-CONF* scores and adding them to the initial tweet.

Finally, we transform the extended tweet to its Indri format, then, we send it to the baseline system to extract, from a provided Wikipedia corpus, a set of sentences representing the tweet context that does not exceed 500 words (this limit is established by the organizers).

[5] it is an open source project downloaded at http://www.cs.rpi.edu/ zaki/www-new/pmwiki.php/Software/Software

5 Experiments, Results and Discussion

In this section, we detail the experimental study of applying our proposed approach, on the issue of tweets contextualization. Experimentations are done using WikipediaMiner, which is a toolkit developed for tapping the rich semantics encoded within Wikipedia, to select the related Wikipedia articles for a given tweet.

This toolkit helps to integrate Wikipedia's knowledge into applications , by:

- Providing simplified, object-oriented access to Wikipedia's structure and content.
- Measuring how terms and concepts in Wikipedia are connected to each other.
- Detecting and disambiguating Wikipedia topics when they are mentioned in documents.

We validated our approach over INEX 2014 [3] collection which contains:

1. A collection of articles, that has been rebuilt based on a dump of the English Wikipedia from November 2012. It is composed of 3 902 346 articles, where all notes and bibliographic references that are difficult to handle are removed and only non-empty Wikipedia pages (pages having at least one section) are kept.
2. A collection of English tweets, composed of 240 tweets selected from the CLEF RepLab 2013. To focus on content analysis alone, urls are removed from the tweets.

We have evaluated our run according to the **Informativeness** metric, this latter is proposed by the INEX organizers, it aims at measuring how well the summary helps a user understand the tweets content. Therefore, for each tweet, each passage will be evaluated independently from the others, even in the same summary. the results are based on a thorough manual run on 1/5 of the 2014 topics using the baseline system. From this run two types of references were extracted, namely:

- a list of relevant sentences per topic.
- extraction of Noun Phrases from these sentences together with the corresponding Wikipedia entry.

The dissimilarity between a reference text and the proposed summary is given by:

$$Dis(T,S) = \sum_{t \in T} (P-1) \times \left(1 - \frac{min(log(P), log(Q))}{max(log(P), log(Q))}\right) \tag{6}$$

where :

- $P = \frac{f_T(t)}{f_T} + 1.$
- $Q = \frac{f_S(t)}{f_S} + 1.$
- T, a set of query terms present in reference summary and for each $t \in T$.

- $f_T(t)$, the frequency of term t in reference summary.
- S, a set of query terms present in a submitted summary and for each $t \in S$.
- $f_S(t)$, the frequency of term t in a submitted summary.

Organizers used three different distributions for the reference summaries in the 2014 tracks, namely:

- Unigrams made of single lemmas (after removing stop-words).
- Bigrams made of pairs of consecutive lemmas (in the same sentence).
- Bigrams with 2-gaps also made of pairs of consecutive lemmas but allowing the insertion between them of a maximum of two lemmas (Also referred to as skip distribution).

We have compared our runs with the following different runs submitted by INEX 2014 participants:

1. In *Best run* participants [21] used association rules between terms to extend the tweets.
2. In *run-Cortex* participants [19] used a statistical summarizer system called Cortex, which is based on the fusion process of several different sentence selection metrics.
3. In *run-Artex* participants [19] used a statistical summarizer system called Artex, which is based on the inner product of main topic and pseudo-words vectors.

We conducted two runs, namely:

- **run-ESA**: in this run, we used the *ESA* to calculate the relatedness between the tweet and the candidates terms.
- **run-ESA-CONF**: in this run, we used our proposed measure to calculate the relatedness between the tweet and the candidates terms, As parameters, CHARM takes *minsupp* = 15 as the relative minimal support and *minconf* = 0.7 as the minimum confidence of the rules. While considering the $Zipf$ distribution of the collection, the minimal threshold of the support value is experimentally set in order to spread trivial terms which occur in the most of the documents, and are then related to too many terms. We realized this run with the best parameter value $\alpha = 0.5$ obtained by experiments.

Tables 1 and 2 describe our obtained results where the lowest scores represent the best runs. This is justified by the fact that the results are diverging. To compare our runs (run-ESA, run-ESA-CONF) with the Best-Run, we perform the Student T-test. The results for Best-run and *run-ESA* seem to be very similar while the run *run-ESA-CONF* has achieved the best informativeness results.

The gray rows of the tables depict the results of our runs, where '††' represents the highly significant results and '†' the ones similar to the Best-run (with p-value <0.05).

Table 1. INEX Tweet Contextualization 2014 informativeness results based on sentences.

Run	Unigrams	Bigrams	Bigrams with 2-gaps
run-ESA-CONF	**0.7613**††	**0.8629**††	**0.8638**††
run-ESA	**0.7665**†	**0.8661**†	**0.8668**†
Best-run	**0.7632**	**0.8689**	**0.8702**
run-Cortex	0.8415	0.9696	0.9702
run-Artex	0.8539	0.9700	0.9712

Table 2. INEX Tweet Contextualization 2014 informativeness results based on noun phrases.

Run	Unigrams	Bigrams	Bigrams with 2-gaps
run-ESA-CONF	**0.7839**††	**0.9229**††	**0.9434**††
run-ESA	**0.7931**†	**0.9269**†	**0.9460**†
Best-run	**0.7903**	**0.9273**	**0.9461**
run-Cortex	0.8477	0.971	0.9751
run-Artex	0.8593	0.9709	0.9752

5.1 Discussion

Our approach achieved promising informativeness results, this is due to the use of Wikipedia that augmented the tweet representation with massive amounts of world knowledge, and to the term ranking phase that reduced the noise in the extended tweets by eliminating the non related terms, and fine-grained the semantic representation of the tweets.
The run *run-ESA-CONF* has achieved the best informativeness results. This can be explained by the use of association rules that led to the enforcement of the relatedness score between the candidate terms and the tweet, ensured that the extended tweets contain adequate correlating terms with the initial ones and helped avoid inclusion of non-similar terms in them as much as possible, so the extended tweets were, to some extent, clean.
At the end, we can say that our proposed approach is able to successfully reduce the tweet drift and allowed us to achieve the goal of this work.

6 Conclusion

In this paper, we proposed a semantic approach for the tweet contextualization task, based on Wikipedia as an external knowledge source. We conducted our experimentations on the INEX 2014 collection. The results we obtained through the different performed runs showed a significant improvement in the informativeness of the contexts. In our future work, we intend to extend our approach to take into account the tweets specificities (#,@,...). Furthermore, we mean to generalize our approach to contextualize normal (regular) queries by applying it on other data collections such as ChiC. This latter contains short queries that

have no sufficient information to express their semantic. Moreover, we plan to use other structured and semantically enriched data sources, such as DBPedia, UMBEL, Freebase, WordNet etc., as external sources.

Acknowledgments. This work is partially supported by the French-Tunisian project PHC-Utique RIMS-FD 14G 1404.

References

1. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C., May 26-28, 1993. pp. 207–216 (1993), http://doi.acm.org/10.1145/170035.170072
2. Bandyopadhyay, A., Ghosh, K., Majumder, P., Mitra, M.: Query expansion for microblog retrieval. IJWS 1(4), 368–380 (2012), http://dx.doi.org/10.1504/IJWS.2012.052535
3. Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., Tannier, X.: Overview of INEX tweet contextualization 2013 track. In: Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013. (2013), http://ceur-ws.org/Vol-1179/CLEF2013wn-INEX-BellotEt2013.pdf
4. Ben-jabeur, L.: Leveraging social relevance: Using social networks to enhance literature access and microblog search. Ph.D. thesis, Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier) (2013)
5. Deveaud, R., Boudin, F.: Contextualisation automatique de tweets à partir de wikipédia. In: CORIA 2013 - Conférence en Recherche d'Infomations et Applications - 10th French Information Retrieval Conference, Neuchâtel, Suisse, April 3-5, 2013. pp. 125–140 (2013)
6. Deveaud, R., Boudin, F.: Effective tweet contextualization with hashtags performance prediction and multi-document summarization. In: Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013. (2013)
7. Ermakova, L., Mothe, J.: IRIT at INEX 2012: Tweet contextualization. In: CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012 (2012), http://ceur-ws.org/Vol-1178/CLEF2012wn-INEX-ErmakovaEt2012.pdf
8. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007. pp. 1606–1611 (2007), http://dli.iiit.ac.in/ijcai/IJCAI-2007/PDF/IJCAI07-259.pdf
9. Gabrilovich, E., Markovitch, S.: Wikipedia-based semantic interpretation for natural language processing. J. Artif. Intell. Res. (JAIR) 34, 443–498 (2009), http://dx.doi.org/10.1613/jair.2669
10. Ganter, B., Wille, R.: Formal concept analysis - mathematical foundations. Springer (1999)
11. Latiri, C.C., Yahia, S.B., jean-pierre chevallet, Jaoua, A.: Query expansion using fuzzy association rules between terms. In: JIM'2003, France, 2003
12. Lau, C.H., Li, Y., Tjondronegoro, D.: Microblog retrieval using topical features and query expansion. In: Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, November 15-18, 2011 (2011), http://trec.nist.gov/pubs/trec20/papers/QUT1.microblog.pdf

13. Linhares, A.C.: An automatic greedy summarization system at INEX 2013 tweet contextualization track. In: Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013. (2013), http://ceur-ws.org/Vol-1179/CLEF2013wn-INEX-CarneiroLinhares2013.pdf
14. Milne, D.N., Witten, I.H.: An open-source toolkit for mining wikipedia. Artif. Intell. 194, 222–239 (2013), http://dx.doi.org/10.1016/j.artint.2012.06.007
15. Morchid, M., Dufour, R., Linéars, G.: Lia@inex2012 : Combinaison de thèmes latents pour la contextualisation de tweets. In: 13e Conférence Francophone sur l'Extraction et la Gestion des Connaissances. Toulouse,France (2013)
16. Shekarpour, S., Höffner, K., Lehmann, J., Auer, S.: Keyword query expansion on linked data using linguistic and semantic features. In: 2013 IEEE Seventh International Conference on Semantic Computing, Irvine, CA, USA, September 16-18, 2013. pp. 191–197 (2013), http://dx.doi.org/10.1109/ICSC.2013.41
17. Song, M., Song, I., Hu, X., Allen, R.B.: Integration of association rules and ontologies for semantic query expansion. Data Knowl. Eng. 63(1), 63–75 (2007)
18. Tan, K.L., Almasri, M., Chevallet, J., Mulhem, P., Berrut, C.: Multimedia information modeling and retrieval (MRIM) /laboratoire d'informatique de grenoble (LIG) at chic2013. In: Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013. (2013), http://ceur-ws.org/Vol-1179/CLEF2013wn-CHiC-TanEt2013.pdf
19. Torres-Moreno, J.: Three statistical summarizers at CLEF-INEX 2013 tweet contextualization track. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. pp. 565–573 (2014), http://ceur-ws.org/Vol-1180/CLEF2014wn-Inex-TorresMoreno2014.pdf
20. Zaki, M., Hsiao, C.J.: An efficient algorithm for closed itemset mining. In: Second SIAM International Conference on Data Mining (2002)
21. Zingla, M.A., Ettaleb, M., Latiri, C.C., Slimani, Y.: INEX2014: tweet contextualization using association rules between terms. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. pp. 574–584 (2014), http://ceur-ws.org/Vol-1180/CLEF2014wn-Inex-ZinglaEt2014.pdf