INEX2014: Tweet Contextualization Using Association Rules between Terms

Meriem Amina Zingla[1], Mohamed Ettaleb[2], Chiraz Latiri[2], and Yahya Slimani[1]

[1] University of Carthage, INSAT, LISI research Laboratory, Tunis, Tunisia
[2] University of Tunis El Manar, Faculty of Sciences of Tunis, LIPAH research Laboratory, Tunis ,Tunisia
{zinglameriem,ettaleb.mohamed1,yahya.slimani}@gmail.com
chiraz.latiri@gnet.tn

Abstract. Tweets are short messages that do not exceed 140 characters. Since they must be written respecting this limitation, a particular vocabulary is used. To make them understandable to a reader, it is therefore necessary to know their context. In this paper, we describe our approach submitted for the tweet contextualization track in CLEF 2014 (Conference and Labs of Evaluation Forums). This approach allows the extension of the tweet's vocabulary by a set of thematically related words using mining association rules between terms.

Keywords: Information retrieval, Tweet contextualization track, Query Expansion, Association rules.

1 Introduction

Web 2.0 is the term associated with the transition of the World Wide Web from a collection of individual web sites to an emerging platform in its own right. This emergence is due largely to users collaborations, these users have been the driving force for the emergence of new services [1]. One of those is the microblogging service, *e.g.*, Twitter, which is a communication medium and a collaboration system that allows broadcasting short messages.

In contrast to traditional blogs, media-sharing and social networks services, microblogs (tweets) are textual messages submitted in real-time to report an idea, an actual interest, or an opinion [2]. The size of these messages may be limited by a maximum number of characters. This constraint, related to the size of message, causes the use of a particular vocabulary. The aim is to exchange a maximum of information in as little characters as possible [3].

In this respect, we will focus on the Tweet Contextualization track. The participants of INEX 2014[1] are required to perform the task of contextualizing tweets, *i.e.*, given a tweet and a related entity, they try to answer questions of the form "why this tweet concerns this entity? should it be an alert?".These questions can be answered by several sentences or by an aggregation of texts from different articles of Wikipedia.

[1] https://inex.mmci.uni-saarland.de/

The purpose of this task is to allow the reader a better understanding of the tweet. This task can be divided into two subtasks. The first is to find the most relevant Wikipedia articles using an IR system, and the second is to extract the most representative passages of the tweet from the relevant Wikipedia articles.

In this paper, we propose to use a statistical method based on association rules mining [4] to extend the tweets. The main thrust in the proposal is that the proposed approach gathers a minimal set of rules allowing an effective selection of rules to be used in the expansion process.

The remainder of this paper is organized as follows: Section 2 cites some related works. Then, Section 3 describes the test data. Section 4 recalls the basic definition for the derivation of association rules between terms and details our proposed approach for tweets contextualization based on these association rules. Next, Section 5 describes our different submitted runs for the tweet contextualization track as the experiment results. The conclusion and future work are finally presented in Section 6.

2 Related works

Though the idea to contextualize tweets is quite recent, there are several works in this context. Recently, Morchid *et al.* [3] used latent Dirichlet analysis (LDA) to obtain a representation of the tweet in a thematic space. This representation allows to find a set of latent topics covered by the tweet. Deveaud *et al.* in [5] used various techniques involving automatic summarization and topic modeling algorithms to score the candidate sentences. Another approach in [6] 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. Cher *et al.* [7] proposed a twitter retrieval framework that focuses on using topical features, combined with query expansion using pseudo-relevance feedback (PRF) to improve microblogs retrieval results. Saad El Din *et al.* in [8] performed a time-bounded web search with the original query to get web results from the same period of the query; then they extracted the web page title of the first web result and used it to extend the original query before applying microblog search. In [9], Ermakova *et al.* proposed a new method based on the local Wikipedia dump. They used TF-IDF [2] cosine similarity measure enriched by smoothing from local context, named entity recognition and part-of-speech weighting presented at INEX 2011. They modified this method by adding bigram similarity, anaphora resolution, hashtag processing and sentence reordering. 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.

[2] TF-IDF : Term Frequency-Inverse Document Frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

3 Description of the Test Data

The tested INEX 2014 collections contains:

1. A collection of articles, that has been rebuilt based on a recent dump of the English Wikipedia from November 2012, it is composed of 3 902 346 articles, 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. Resulting documents are made of a title (title), an abstract (a) and sections (s). Each section has a sub-title (h). Abstract and sections are made of paragraphs (p) and each paragraph can have entities (t) that refer to Wikipedia pages. Each document is provided in XML format and respects the Document Type Definition (DTD) described in Table 1.

Table 1. DTD of Wikipedia pages

<!ELEMENT xml (page)+>
<!ELEMENT page (ID, title, a, s*)>
<!ELEMENT ID (#PCDATA)>
<!ELEMENT title (#PCDATA)>
<!ELEMENT a (p+)>
<!ELEMENT s (h, p+)>
<!ATTLIST s o CDATA #REQUIRED>
<!ELEMENT h (#PCDATA)>
<!ELEMENT p (#PCDATA — t)*>
<!ATTLIST p o CDATA #REQUIRED>
<!ELEMENT t (#PCDATA)>
<!ATTLIST t e CDATA #IMPLIED>

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. An example of a tweet representation is given in Table 2. Each tweet is composed of:

   - An identifier (Id),
   - A category, the tweet collection is divide into four (4) distinct categories: automotive, music, banking and university.
   - An entity, there are 64 distinct entities name from Wikipedia.
   - A topic, there are 235 distinct manual topic labels.
   - A text content, which must not exceed 140 characters.
   - A timestamp.

Table 2. Example of how tweets are represented in the INEX 2014 collections

```
<tweet id=”260979833180397568”>
< category> automotive </category>
<entity> AB Volvo </entity >
< topic> company sales report </topic>
< content> Bad news from the world’s no. 2 truck maker... Volvo Q3 operating profit ...
4.5B last year. sees tough quarter ahead.</content>
<timestamp> 2012-10-24-07:43</timestamp>
</tweet>
```

4 A novel approach for tweets contextualization based on association rules between terms

4.1 Association rules mining and Information Retrieval

Given a user query, the number of retrieved documents can be overwhelmingly large, hampering their efficient exploitation by the user. In this situation, the query expansion technique offers an interesting solution for obtaining a complete answer while preserving the quality of retained documents. In [10], authors presented a novel approach for mining knowledge supporting query expansion that is based on association rules [4]. The key feature of this approach is a better trade-off between the size of the mining result and the conveyed knowledge. An experimental study has examined the application of association rules to some real collections, whereby automatic query expansion has been performed. The results showed a significant improvement in the performances of the information retrieval system, both in terms of recall and precision.

An association rule R is an implication of the form R: $T_1 \Rightarrow T_2$, where T_1 and T_2 are subsets of $\mathcal{T}$, 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 as 1,2:

$$Supp(R) = Supp(T), \tag{1}$$

while its *confidence* is computed as:

$$Conf(R) = \frac{Supp(T)}{Supp(T_1)}. \tag{2}$$

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

Interestingly enough, to address tweet contextualization in an efficient and effective manner, we claim that a synergy with some advanced text mining methods, especially association rules [11], is particularly appropriate. However, applying association rules in the context of tweet contextualization is far from being a trivial task, mostly because of the huge number of potentially interesting rules that can be drawn from a document collection.

4.2 Proposed approach

The tweet contextualization system serves to expand a given tweet and to elaborate the corresponding query, which is sent in order to retrieve its related context. Our proposed approach based on association rules between terms is depicted in Figure 1. The contextualization tweet process is performed on the following steps:

1. Selecting of a sub-set of articles, according to the tweet's subject, from the documents collection.
2. Annotating the selected Wikipedia articles with part-of-speech and lemma information using TreeTagger [3].
3. Extracting of only nouns from the annotated Wikipedia articles, and removing the most frequents nouns.
4. Generating the association rules using an efficient algorithm CHARM [4] for mining all the closed frequent termsets [12]. As an input, CHARM takes a corpus in the basic ascii format, where each line or row (article) is a list of terms, *minsupp* as the relative minimal support and $minconf$ as the minimum confidence of the rules, and gives as output, the association rules with their appropriate support and confidence (*cf.* Figure2) [12] .
5. Projecting the tweets on the set of the association rules in order to obtain the thematic space of each tweet. This is done by projecting the terms of the tweet on the premises of the association rules and enriching the tweet using their conclusions.
6. Creating the query from the terms of the tweet and the thematic space, this query is then transformed to its Indri [5] format.
7. Sending the query to the baseline system, composed of an Information Retrieval System (IRS) and an Automatic Summary System (ASS) offered by the organizers of INEX 2014, to extract from a provided Wikipedia corpus a set of sentences representing the tweet context that should not exceed 500 words.

[3] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

[4] CHARM: Closed Association Rule Mining; the H is gratuitous, it is an open source downloaded at http://www.cs.rpi.edu/ zaki/www-new/pmwiki.php/Software/Software

[5] Indri is a search engine that provides state-of-the-art text search and a rich structured query language for text collections of up to 50 million documents (single machine) or 500 million documents (distributed search). Available for Linux, Solaris, Windows and Mac OSX. http://www.lemurproject.org/indri.php

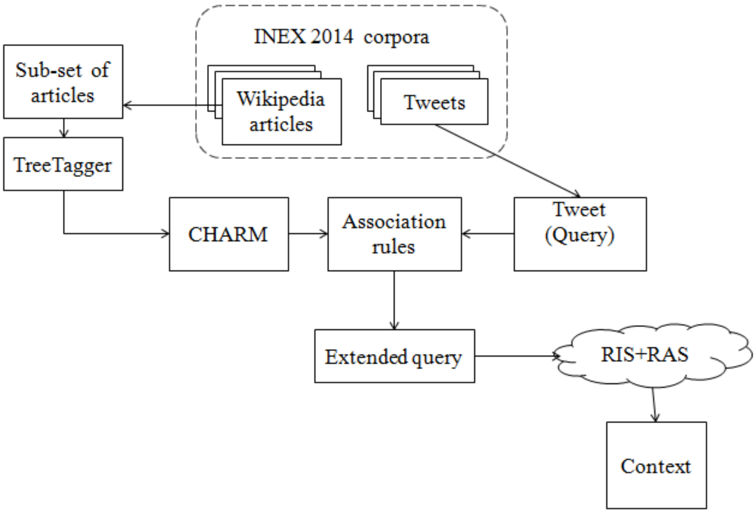
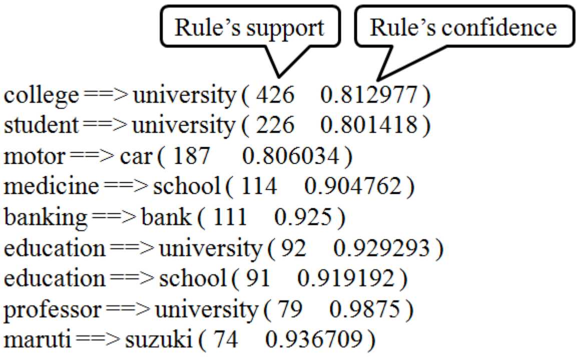


Fig. 1. The Proposed approach steps


Rule's support
Rule's confidence

```
college ==> university ( 426   0.812977 )
student ==> university ( 226   0.801418 )
motor ==> car ( 187   0.806034 )
medicine ==> school ( 114   0.904762 )
banking ==> bank ( 111   0.925 )
education ==> university ( 92   0.929293 )
education ==> school ( 91   0.919192 )
professor ==> university ( 79   0.9875 )
maruti ==> suzuki ( 74   0.936709 )
```

Fig. 2. Example of generated rules

An illustrative example is depicted in Figure 3.

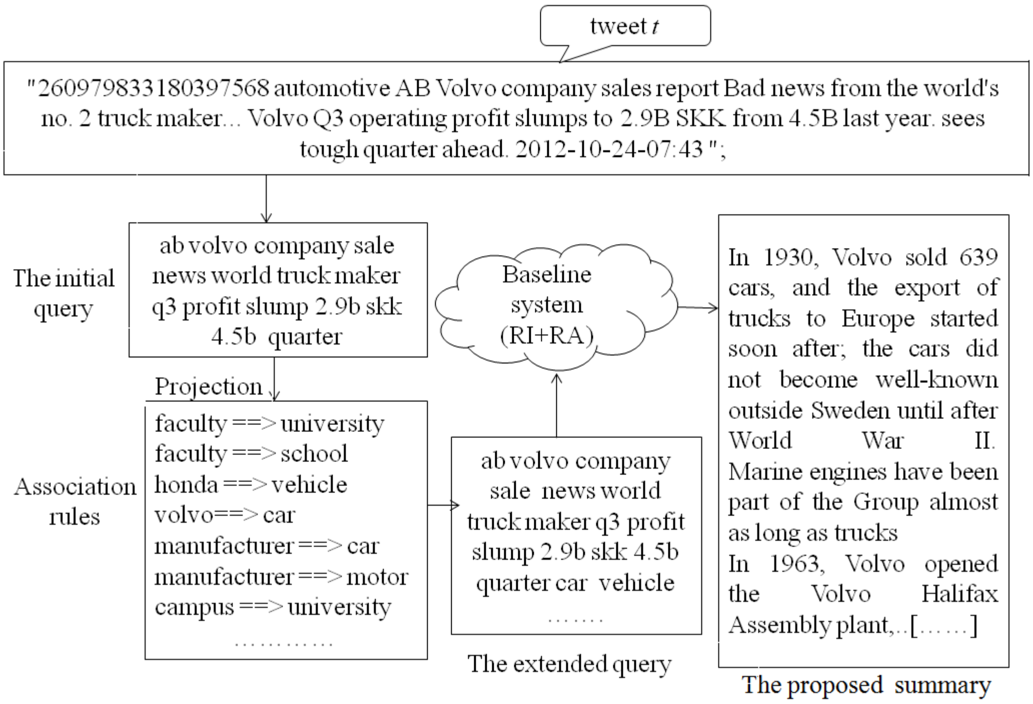

tweet *t*
"260979833180397568 automotive AB Volvo company sales report Bad news from the world's no. 2 truck maker... Volvo Q3 operating profit slumps to 2.9B SKK from 4.5B last year. sees tough quarter ahead. 2012-10-24-07:43 ";
The initial query
ab volvo company sale news world truck maker q3 profit slump 2.9b skk 4.5b quarter
Projection
Association rules
faculty ==> university
faculty ==> school
honda ==> vehicle
volvo==> car
manufacturer ==> car
manufacturer ==> motor
campus ==> university
....
Baseline system (RI+RA)
ab volvo company sale news world truck maker q3 profit slump 2.9b skk 4.5b quarter car vehicle
....
The extended query
In 1930, Volvo sold 639 cars, and the export of trucks to Europe started soon after; the cars did not become well-known outside Sweden until after World War II. Marine engines have been part of the Group almost as long as trucks
In 1963, Volvo opened the Volvo Halifax Assembly plant,..[……]
The proposed summary

Fig. 3. Example of how to contextualize a tweet using association rules between terms

5 Experiments and results

5.1 Evaluation Framework

We released three runs that are based on the approach described in Section 4.2, namely :

Run1 (ID = 359): We selected a sub-set of 50 000 articles, according to the entities associated to tweets, using an algorithm that calculates the similarity between the entities associated to tweets and the articles from the documents collection. Then, we proceeded as in the proposed approach by applying the algorithm CHARM with the following parameters : $minsupp = 16$, and $minconf = 0.7$.

Run2 (ID = 360): In the second run, we used the entities associated to tweets to extract their context, then we exploited these contexts considering them as articles, we proceeded, afterwards, as in the proposed approach. We applied the CHARM algorithm taking $minsupp = 5$ and $minconf = 0.7$ as parameters.

Run3 (ID = 359): In this run, we used the recommended documents list for query, provided by INEX. As in runs 1 and 2, we proceeded, afterwards, as in the proposed approach. We applied the CHARM algorithm taking $minsupp = 5$ and $minconf = 0.7$ as parameters.

5.2 Evaluation metrics

All of these runs are evaluated according to [13]:

Informativeness: Informativeness aims at measuring how well the summary helps a user to understand the tweet 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.

Two reference runs using 2013 and 2012 corpus were also generated to see if the evolution of the Wikipedia had an impact.

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

$$Dis(T,S) = \sum_{t \in T} (P-1) \times \left(1 - \frac{min(log(P), log(Q))}{max(log(P), log(Q))}\right) \tag{3}$$

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).

INEX evaluated 12 runs from four participants. Our result runs (IDs are 361, 360 and 359) ranked first, second and fifth, respectively, in the informativeness evaluation based on sentences (*cf.* Table3). Table 4 highlights informativeness quality -based on noun phrases- difference between the original queries (Ref Runs) and the expanded ones using association rules mined with the algorithm CHARM. Our result runs 361, 360 and 359 are ranked first, fourth and sixth, respectively.

Table 3. Informativeness based on sentences

Run Id	Rank	Unigram	Bigram	Skip
ref2013	1	0.705	0.794	0.796
ref2012	2	0.7528	0.8499	0.8516
361	**3**	**0.7632**	**0.8689**	**0.8702**
360	**4**	**0.782**	**0.8925**	**0.8934**
359	**7**	**0.8022**	**0.912**	**0.9127**

Table 4. Informativeness based on noun phrases

Run Id	Rank	Unigram	Bigram	Skip
ref2013	1	0.7468	0.8936	0.9237
ref2012	2	0.7784	0.917	0.9393
361	**3**	**0.7903**	**0.9273**	**0.9461**
360	**6**	**0.8104**	**0.9406**	**0.9553**
359	**8**	**0.8227**	**0.9487**	**0.9613**

Readability: readability aims at measuring how clear and easy it is to understand the summary. Each summary has been evaluated by considering the following four parameters:

- Readable: Percentage of passages considered as readable (non trash).
- Syntax: Percentage of passages without syntax or grammatical errors.
- Diversity: Percentage of non redundant passages.
- Structure: Percentage of non breaking anaphora passages.

Evaluation was carried out by one reader over a pool of 12 summaries per run. Our runs, 359,360 and 361, ranked eighth, fifth and eleventh respectively on the official evaluation. Table 5) shows the runs from the 2014 readability evaluations.

The obtained informativeness evaluation results shed light that our proposed approach, based on association rules, offers interesting results and helps ensure that context summaries contain adequate correlating information with the tweets and avoid inclusion of non-similar information in them as much as possible.

Table 5. Readability evaluation

Run Id	Rank	Readable	Syntax	Diversity	Structure	Avg
360	**5**	**92.6%**	**70.35%**	**58.84%**	**86.33%**	**77.03%**
ref2013	6	91.74%	69.82%	60.52%	85.80%	76.97%
ref2012	7	91.39%	69.58%	60.67%	85.56%	76.80%
359	**8**	**93.03%**	**70.64%**	**53.53%**	**86.34%**	**75.88%**
361	**11**	**93.23%**	**70.41%**	**50.12%**	**85.97%**	**74.93%**

Furthermore, our tweets contextualization approach based on association rules leads to the improvement of the context informativeness ; we note also that the selection of articles impacts the quality of the contexts. The contexts are less readable, it may be that they contain some unexpected noises which need to be cleaner.

6 Conclusion

In this paper, we have described a new method for tweet contextualization based on association rules between sets of terms. The experimental study was conducted on INEX 2014 collections. The results confirmed that the synergy between association rules and query expansion is fruitful. Further work investigates whether the combination of the confidence measure with other measures in the weighting process can be of benefit to the tweet contextualization process.

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

References

1. The emergence and empowerment of web 2.0, http://www.dialogic.com/~/media/products/docs/whitepapers/11339-web2-0-wp.pdf.
2. Ben Jabeur, L.,Tamine, L.,Boughanem, M.: Uprising microblogs: a bayesian network retrieval model for tweet search. In Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC '12). ACM, New York, NY, USA, 943-948,(2012).
3. Morchid, M., Linarès, G.: INEX 2012 Benchmark a Semantic Space for Tweets Contextualization. In CLEF 2012 Evaluation Labs and Workshop Online Working Notes, Rome, Italy, September 17-20, (2012).
4. Agrawal,R., Imielinski,T., Swami,A.: Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 207-216, Washington D.C., May (1993).
5. Deveaud,R.,Boudin, F.: LIA/LINA at the INEX 2012 Tweet Contextualization track. In: Proceedings of the CLEF 2012 Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2012), Rome (Italy), September 17-20, (2012).

6. Ayan,B., Mandar,M., Prasenjit, M.,: Query expansion for microblog retrieval: International Journal of Web Science, vol.1, pp. 368-380, (2012).
7. Lau, Cher Han and Li, Yuefeng and Tjondronegoro, Dian. Microblog Retrieval Using Topical Features and Query Expansion. . TREC. editor(s) Voorhees, Ellen M. and Buckland, Lori P.. National Institute of Standards and Technology (NIST), (2011).
8. Saad El Din,A., Magdy, W.: Microblog Retrieval Using Topical Features and Query Expansion,(2012).
9. Ermakova, L., Mothe, J.: IRIT at INEX2012: Tweet contextualization In CLEF 2012 Evaluation Labs and Workshop Online Working Notes, Rome, Italy, September 17-20, (2012).
10. Latiri, C., Haddad, H., Hamrouni, T.: Towards an effective automatic query expansion process using an association rule mining approach. In J. Intell. Inf. Syst., (39) 1: 209-247, (2012).
11. Kartick, C.M., Nicolas, P., Anirban, M., Ujjwal, M., Sanghamitra B.: A New Approach for Association Rule Mining and Bi-clustering Using Formal Concept Analysis, (2012).
12. Zaki,M., Hsiao,J.:An efficient algorithm for closed itemset mining :In Proceedings of the Second SIAM International Conference on Data Mining, (2002).
13. Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., Tannier, X.: Overview of INEX Tweet Contextualization 2013 track, (2013).