Geographical Queries Reformulation using Parallel FP-Growth for Spatial Taxonomies Building

Omar El Midaoui[1,2], Btihal El Ghali[2] and Abderrahim El Qadi[3]
[1]LRIT Associated Unit to the CNRST - URAC n°29, Faculty of Sciences, Mohammed V University in Rabat, Morocco
[2]SmartiLab, Ecole Marocaine des Sciences de l'Ingénieur (EMSI), Rabat, Morocco
[3]TIM, High School of Technology, Mohammed V University in Rabat, Morocco

Keywords: Information Retrieval, Parallel FP-Growth Algorithm, Machine Learning, Geographical Query Reformulation, Spatial Entity, Spark, Big Data.

Abstract: Due to its specificities and hierarchical structure, a geographical query needs a special process of reformulation by Information Retrieval Systems (IRS). This fact is ignored by most of web search engines. In this paper, we propose an automatic approach for building a spatial taxonomy that models' the notion of adjacency that can be uses in the reformulation of the spatial part of a geographical query. This approach exploits the documents that are in the top of the list of retrieved results when submitting a spatial entity, which is composed of a spatial relation and a noun of a city. Then, a transactional database is constructed, considering each document extracted as a transaction that contains the nouns of the cities sharing the country of the submitted query's city. The algorithm FP-Growth is applied to this database in his parallel version (PFP) in order to generate association rules, that will form the country's taxonomy in a Big Data context. Experiments has been conducted on Spark and their results show that query reformulation based on the taxonomy constructed using our proposed approach improves the precision and the effectiveness of the IRS.

1 INTRODUCTION

Most human activities are well located in the geographical area. Thus, it is not surprising that a big amount of web documents contain geographical references. A study that was done on the Excite search engine shows that between every five queries there is one query which have a geographical context (Sanderson and Kohler, 2004). Web users searching for information that are spatially located often require information, that are geographically specific, such geographic terms in Web pages and user queries or even user location (Jiang et al., 2018). However, retrieval systems currently have limited support to operationalize a user's geospatial queries. Geographic information deals with physical objects that are in some cases hard to express with words and that contain most of the time ambiguous terms. These arguments prove the fact that it will be very useful for search engines to take into account the spatial scope of geographical queries.

The current search engines generally handle queries by adopting a keyword matching approach without inferring the geographical scope of the spatial terms. Thus, when the name of a place is typed into a typical search engine associated with a spatial preposition (e.g. "near"), web pages that include that name in the text will be retrieved but most likely, not places that are close to that specified place.

In order to do a spatial analysis of text, the first step is the annotation of spatial named entities. Several techniques have proved their ability for carrying out this annotation, such as the works of (Rocío and Erick, 2010) and (Loustau, 2008) that has elaborated it using external resources named "gazetteers". A gazetteer is a dictionary or geographic directory whose inputs are names of places. Each entry in the dictionary may be associated with information such as belonging to one or more administrative structures (town, region, country, etc.), the physical characteristic (mountain, river, road, etc.), statistical data, and a geometric representation expressed in a geographic referential.

Other works proposes the categorization of these spatial named entities after identification. Such as, Buscaldi and Rosso that proposed a technique for spatial named entities categorization using the thesaurus Geo-WordNet (Buscaldi and Rosso, 2008).

El Midaoui, O., El Ghali, B. and El Qadi, A.
Geographical Queries Reformulation using Parallel FP-Growth for Spatial Taxonomies Building.
DOI: 10.5220/0009446603750381
In *Proceedings of the 5th International Conference on Internet of Things, Big Data and Security (IoTBDS 2020)*, pages 375-381
ISBN: 978-989-758-426-8

Or Bouamor that exploited a document structure in (Bouamor, 2009), using as a corpus the collaborative encyclopaedia "Wikipedia". In his work, the identification of named entities is done using the title and their categorization is based on the analysis of the first sentence of the description part or the category part at end of the article.

Different external resources have been also created and used in the most recent works. Such as the Geographical Information Retrieval model, proposed by (Fang and Zhang, 2018), that simplifies the process of user information acquisitions by analyzing and extracting the attribute features, spatial features and structure features of Geographic Markup Language (GML) data. As defined by this work, GML data is generated by artificial of special services and stored semi-structured data using the Geographic Mark-up Language (GML), which is a geographic information coding specification that was implemented based on the Extensible Mark-up Language (XML) and standardized (ISO 19136-2007).

In the other hand, some approaches aim more particularly to the disambiguation of recognized places names (Buscaldi, 2009). The ambiguity can be understood as a word or a phrase that has many meanings (Vargas et al., 2012). In this case, two types of ambiguities are to treat (Einat et al., 2004): a geo/non-geo ambiguity is when the entity has a non-geographical meaning like the term "Turkey", and a geo/geo ambiguity that occurs when the named entity refers to two different places as Rabat in Malta and Rabat in Morocco.

A hybrid approach is proposed in (Gaio et al., 2012) which, first landmark names of places but also searches for these terms in ontological resources to identify related terms, potentially geographic. Domain-Specific taxonomies (Yangqiu et al., 2015) are also playing an important role in many applications for improving search results (Xueqing et al., 2012) or helping with query reformulation as in (Sadikov et al., 2010) and (Aloteibi, Mark Sanderson, 2014).

In this paper, we propose a geographical taxonomy builder using the parallel FP-Growth Algorithm (PFP) which inputs are text documents and we complete the process by suggesting a query reformulation approach for geographical queries. Our contributions are also made in the step of geographical and thematic entities separation giving a geographical query, in order to reformulate the two entities in a different manner.

This approach is tested using a collection that has been created during our experimentations. This collection contains 50 queries and 2500 documents. We used 1500 documents, considering 30 retrieved documents per city for the taxonomy building step, as we used a list of the 50 most popular cities. In addition of, 20 retrieved documents per submitted query in the reformulation step evaluation (10 before and 10 after reformulation). Thus, 2500 documents have been used in total. The collection's documents were retrieved automatically using the google web services whenever there was a need.

This article is organized as follows:

The section 2 is introducing our proposed approach for the construction of a geographical taxonomy of adjacency using the PFP algorithm, while we explain our query reformulation technique in section 3. The results of our experimentations are presented in section 4. Finally, section 5 draws conclusions and future works.

2 THE GEOGRAPHICAL TAXONOMY BUILDER

A taxonomy consists of a number of names arranged in a hierarchical system that describe a specific domain (Enghoff, 2009) by a hierarchical structure. A taxonomy starts from a general concept of a domain, and associate to it the terms that describe this specific domain more precisely while moving down in the hierarchy.

In this work, we introduce an automatic approach that builds a geographical taxonomy of adjacency. In this aim, we exploit the best-ranked documents retrieved using the search engine when submitting a spatial part of a query, that contain the spatial relation of adjacency and a noun of a city for which we are constructing the taxonomy.

The proposed approach is based on the Parallel FP-Growth algorithm. The geographical query model used in this work considers two type of entities: the Absolute and the Relative Spatial Entities (ASE and RSE). The geographical named entities such as the city of "London" are well-known named entities and are defined as an ASE (Absolute Spatial Entity). While complex spatial entities as "near London" are labelled as an RSE (Relative Spatial Entities).

2.1 The FP-Growth Algorithm

The FP-growth technique (Han et al., 2004) is an association rules Machine Learning algorithm, where "FP" is the acronym of Frequent Pattern. Given as

input a dataset of transactions, the first step of this algorithm is to compute item frequencies and identify the most frequent items. Different from Apriori algorithm (Najadat et al., 2013) (Al-Maolegi and Arkok, 2014) designed for the same aim, by its second step that uses a suffix tree structure, called FP-tree, to encode transactions without the explicit generation of candidate sets, which are usually expensive to generate. After this step, the frequent item sets are extracted from the FP-trees.

The FP-growth is a two phases algorithm. The first phase consists on the construction of FP-Trees and the second mines frequent patterns from the generated FP-Trees.

The construction of an FP-Tree requires two scans on the used database. The first scan permits the selection of the frequent items that are then sorted based on their frequency in descending order to form a new structure caller F-list. The second scan constructs the FP-Tree. First, the non-frequent items are removed while reordering the database tuples according to F-list. Then the reordered transactions are inserted into the FP-Tree. The Input of the Growth part of the algorithm is the constructed FP-Tree and the value of minimum support threshold.

FP-Growth traverses' nodes in the FP-Tree beginning from the least frequent item in F-list. While traversing each node, FP-Growth collects items on the path from the node to the root of the tree. Those collected items constitute the elements of the conditional pattern base of the current item in F-list. The conditional pattern base of an item is defined as a small database of patterns that co-occur with this item. Then FP-Growth creates small FP-Trees based on the conditional pattern bases and re-executes the algorithm recursively on the new FP-Trees until no conditional pattern base can be generated.

2.2 The Parallel FP-Growth

The parallelized FP-growth work on distributed machines (Lingling and Yuansheng, 2015). Its partitions computation is done in such a way that each machine executes an independent group of mining tasks. This method of partitioning eliminates computational dependencies between machines, and thereby communication between them.

Given a transaction database DB, the PFP algorithm's steps are as follows:

- *Sharding*: splitting DB into successive parts and storing those parts on n different machines. Each resultant part is called a shard.
- *Parallel Counting*: counting the support values of all items appearing in each shard. This step permits to discover the items' vocabulary implicitly, which is normally unknown for a huge Database. The result of this step is an F-list.
- *Grouping Items*: Considering I the set of vocabulary discovered, splitting the |I| items appearing in F-List into Q groups. The groups list is called G-list, where each group is given a unique group-id (gid). As F-list and G-list are both small, this step can be executed on a single node of the cluster in few seconds.
- *Parallelizing*: Selecting group-dependent transactions on which the FP-Growth algorithm is applied in order to build local FP-trees in parallel and growth their conditional FP-trees recursively.
- *Aggregating*: Aggregating the results generated in Step 4 as our final result.

PFP distributes the growing FP-trees work based on the transactions' group. thus, this approach is more scalable than a single-node implementation.

PFP is implemented in MLlib on Spark and it takes three parameters: the minimum support threshold to identify frequent item sets, the minimum confidence for generating Association Rules and the number of shards used to distribute the job.

2.3 Geographical Taxonomy of Adjacency

Considering a database whose transactions are documents and items are the cities of the country that contain the city of the user query. We propose to build a spatial taxonomy (Fig. 1) of adjacency based on the PFP algorithm.

The documents that form the input transactional database are restricted to the Absolute Spatial entities contained in the documents. Thus, the items considered are the ASEs.

After the application of the PFP algorithm, starting from the capital of the country for which we will built the taxonomy, the fusion of all the generated FP-trees is forming our geographical taxonomy.

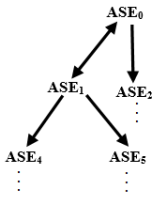

Figure 1: A two-level taxonomy for the ASE0.

Validation Step. In this contribution, we propose also a step of validation of each arc of the taxonomy. To validate each arc in the aggregating step, we verify if its two parts (the two ASEs that form this arc) mutually generate each other in the FP-trees. For example, ASE_1 involves ASE_0 and ASE_0 involves ASE_1 then the arc is kept and this taxonomy evolves to a two-level taxonomy as shown in figure 1. Otherwise, ASE_0 has involved ASE_3, but ASE_3 did not involve ASE_0 so this arc has not been validated. Thus, it has been removed from the taxonomy.

3 REFORMULATION OF A GEOGRAPHIC QUERY USING A TAXONOMY

In order to reformulate a geographical query, we first separate the different components of the query based on the approach of geographic information extraction (GIE) proposed in (Sallaberry et al., 2007). This approach utilizes a methodology of semantic annotation for the detection of geographical markers: first, the Absolute Spatial entity is detected and annotated. Then the spatial entity (SE) is constructed considering this ASE and a lexicon of spatial relations. The remaining words of the query form the thematic entity (TE).

In this step also, we proposed a contribution. We made some modifications in the GIE approach cited above based on a hypothesis.

Hypothesis. If the Spatial Relation is not present in the query, the occurrence of an ASE does not mean that the query has a geographical intent. For example, a query containing “George Washington”. We can also consider the example of the query searching for “Hôtel de Paris”. In this context, the noun “Paris” is the name of a hotel whose location is in Tangier, Monte Carlo or Monaco.

After the separation of the different entities of the user query, we continue applying the proposed approach by interpreting the spatial relationship contained in the spatial entity of the query. The interpretation is done using a lexicon of adjacency spatial relations. The process of reformulation depends on the result of this interpretation.

If the spatial relation detected in the query is an SR of adjacency, we reformulate the spatial part of the query using the country’s taxonomy (Xueqing et al., 2012).

Logically, a query that contains a relation of adjacency means that the intent of the user is to retrieve places that are around the ASE of his query. Thus, we propose to eliminate the entire spatial entity, and to replace by the direct child-nodes items (CNIs) of the query’s ASE in the geographical taxonomy as follows:

- *User New Query = TE SR ASE*
- *Reformulated Query = expanded TE + “CNI 1” or “CNI 2” or …*

In the query resulted from the reformulation, quotes are used to search for the desired place and not separately search for the words that the place’s name contains if the ASE is composed many terms (e.g. the submission of New York unquoted, can lead the search engine to search for New and York as two independent terms). Moreover, the boolean operator ‘or’ is used, to ensure that the retrieval returns documents that include for example "CNI 1" or "CNI 2" or both of them and so on for all the child-nodes used to reformulate the query.

4 EXPERIMENTATION RESULTS

To apply the proposed approach, we have used a lexicon of spatial relationships, and a database of validated ASEs associated with their countries.

In order to test and verify the performance of the technique of taxonomy building proposed in this work, we take our country Morocco as an example. Thus, to be able to use the web pages created by Moroccans themselves we perform our tests in French. “Rabat”, the capital of Morocco is the ASE that we took as a root of our taxonomy. The search engine used in our experimentations is Google web service.

We apply our method using transaction database that is constructed by iterating on Morocco’s ASEs list (a list of 50 cities and villages of Morocco). For every ASE, we selected the thirty first web pages retrieved when submitting a RSE containing the current ASE.

As a pre-treatment step, we deleted accents from the extracted documents to minimize the matching gab between ASEs, due to different manners write cities names by the persons who wrote the documents contents. Because, the miss-matching problem arise particularly in the case of nouns, which contain accents.

Then, we varied the SR of the spatial entities submitted to verify if the variation of the SR influences the performance of the proposed approach. The spatial relations used in this test step are as follows:

Table 1: Spatial Relations.

Annotation	Expression
SR 1	à côté de
SR 2	à la périphérie de
SR 3	à proximité de
SR 4	aux alentours de
SR 5	aux environs de
SR 6	les environs de
SR 7	près de

First, the five top-ranked documents were extracted for the ASE Rabat associated with every spatial relationship of Table 1. A database (DB) containing 35 transactions is constructed based on these documents. The parallel FP-growth algorithm is applied to this DB and then the association rules are generated between Rabat and every Moroccan ASE that co-occur with it in the DB.Ater that, we varied the minsup from 0,2 to 0.8 (as shown in Table 2) without the validation step for the rules extracted using 2-frequent item sets. Later we computed the error rate and the number of rules generated in every case.

Table 2: The error rate and the number of rules generated while varying the minimum support threshold and the spatial relationship used for item sets containing the ASE "Rabat".

	Error rate				Number of generated rules			
RS\ Threshold	0,2	0,4	0,6	0,8	0,2	0,4	0,6	0,8
RS 1	72,73	28,57	33,33	0	22	7	3	1
RS 2	42	0	0	0	5	2	1	1
RS 3	25	0	0	-	4	1	1	0
RS 4	40	33,33	0	0	10	3	1	1
RS 5	33,33	0	0	0	9	2	2	1
RS 6	40	50	0	0	10	4	2	2
RS 7	0	0	0	-	6	6	1	0

From Table 2, we notice that using the minsup=0,8 the algorithm does not return any results in some cases otherwise it gives 1 or 2 answers. The same for minsup=0,6 that do not exceed 2 correct answers.

Regarding the value 0,2 it generally gives a high error rate and sometimes returns a very high number of responses up to 22 resulting ASEs in the case of RS 1 with 6 correct adjacent ASEs only. Thus, we favored the value of minimum support equal to 0,4 because it is the one that gives the best ratio between a minimal error and an acceptable number of answers.

The next step of experimentations is done in order to compare the cases where we use or not the validation step for aggregating the generated FP-trees in order to built the taxonomy of adjacency, based on a minimum support of 0,4.

Table 3: The error rate and the number of correct rules generated using the step of validation or not and using the average of support between the two cases, varying the spatial relation used for item sets containing the ASE "Rabat" with a minsup of 0,4.

	Error rate			Number of correct rules		
Spatial relation	WV	UV	AS	WV	UV	AS
SR 1	28,57	0	0	7	2	2
SR 2	0	0	0	2	1	1
SR 3	0	-	0	1	0	1
SR 4	33,33	50	33,33	3	2	3
SR 5	0	0	0	2	2	2
SR 6	50	0	0	4	2	2
SR 7	0	0	0	6	5	6

WV: without validation, UV: Using validation, AS: Average support

Comparing the results using validation with the results without validation, we note that the error rate decreases when using the validation step, with the exception of the SR 4 for which from 3 results including 2 correct ASEs, validation has eliminated one of the correct ASEs and kept the erroneous one. Concerning the SR 3 we notice that the only ASE that was resulted without validation was eliminated with the step of validation. In general, we conclude that the validation step reduces errors sufficiently.

To minimize the error rate while keeping as much as possible of correct results (eliminate only the erroneous ASEs by the validation step). We propose to compute the average of the two supports of the opposite rules (e.g. ASE1 → ASE2 and ASE2 → ASE1). Table 4 shows that the result given by the case of the average support solves the problems mentioned above for the SR 3 and SR 4.

Comparing the seven spatial relations, we promote the SR 7 "près de" which gives the most interesting result with 0% error and six correct ASEs as child nodes of Rabat's taxonomy of adjacency.

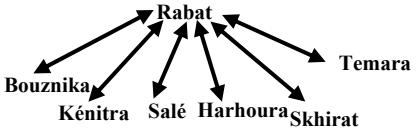

Figure 2: A one-level taxonomy for Rabat using the spatial relation "Près de".

Using the favorable conditions represented above we continue the construction of Morocco's taxonomy (as shown in figure 3) with 0,4 as a minimum support, and using the average of support for validating links.

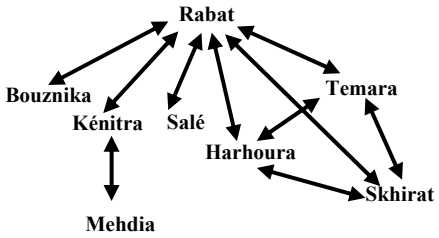

Figure 3: A two-level geographical taxonomy of adjacency for Morocco.

For building this taxonomy, we had used 50 Moroccans ASEs and we were searching for the 30 first retrieved documents while submitting every ASEs with the selected relationship. Thus, 1500 documents have been used in this test with a minimum support of 0,4 and a minimum confident of 0,6. This test have been conducted in 0,85 seconds due to the use of Spark, using a cluster of two nodes.

In order to evaluate the precision of the results of our approach and confirm the results of the precedent tests, we proposed 50 geographical queries that has been submitted to the Google search services with and without reformulation using the taxonomy realized based on the PFP algorithm, and we compared the values of the P@10 and the Mean Average Precision of the two cases.

Table 4: The performance rate of the presented technique according to the original queries.

Baseline	P@10	MAP
Original queries	10,56%	12,92%

From Table 4, we notice that the approach presented in this manuscript gives an interesting improvement in the precision of the geographical queries used in our experiments.

5 CONCLUSION AND FUTURE WORKS

In this paper, we proposed a new method of construction of geographical taxonomies of adjacency using the parallel FP-Growth, and a technique for reformulating geographical queries that contain a spatial entity of adjacency. We have conducted tests on the taxonomy builder method by forming Morocco's taxonomy of adjacency. During our experimentations, we varied the minimum support threshold and the used spatial relationship in order to search for the parameters of the approach that extract the most appropriate frequent item sets. Then we constructed the Moroccan taxonomy using a minimum support of 0.4 and a minimum confidence of 0.6 and the SR number 7, because these conditions gave the best results during our experiments. The proposed technique of reformulation has been tested on 50 queries, which have a geographical intent and their thematic entities are from different fields. These queries had been reformulated based on the spatial taxonomy of adjacency. Finally, we compared the results retrieved by the search engine before and after the application of our technique using the evaluation measures MAP and P@10. The results show that the reformulation based on our proposed approach and using a small number of reformulation terms has improved the value of MAP significantly. Considering the experimental results, we conclude that the presented method is an efficient work that permit to interpret and improve the results of queries containing a spatial entity of adjacency.

As future work, we intend to propose a new method of geographical query reformulation, based on Big Data technologies and an in-depth analysis of user's behaviours through a study of a search engine's trace.

REFERENCES

Al-Maolegi, M., Arkok, B., 2014. *An improved Apriori algorithm for association rules*, International Journal on Natural Language Computing, Vol. 3, Issue 1, pp. 21-29, February 21-29.

Aloteibi, S., Sanderson, M., 2014. *Analyzing geographic query reformulation: An exploratory study*, Journal of the Association for Information Science and Technology, Vol. 65, Issue 1 January, pp. 13-24.

Bouamor, H., 2009. *Extraction des connaissances à partir du Web pour la recherche des images géoréférencées*, in Proceedings of Conférence en Recherche d'Information et Applications (CORIA), pp. 519-526.

Buscaldi, D., 2009. *Toponym ambiguity in geographical information retrieval*, in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '09, ACM, New York USA, pp. 847-847.

Buscaldi, D., Rosso, P., 2008. *Using GeoWordNet for Geographical Information Retrieval*, In proceedings of CLEF, pp. 863-866.

Einat, A., Har'el, N., Sivan, R., Soffer, A., 2004. *Web-a-where: Geotagging Web Content*, Proceedings of the 27th annual international ACM SIGIR conference on

Research and development in information retrieval, Sheffield United Kingdom, pp. 273-280.

Enghoff, H., 2009. *What is taxonomy? – An overview with myriapodological examples*, Soil organisms, Vol. 81, Issue 3, pp. 441–451.

Fang, C., Zhang, S., 2018. *Geographic Information Retrieval Method for Geography Mark-Up Language Data*, ISPRS International Journal of Geo-Information Vol 7, Issue 3.

Gaio, M., Nguyen, V.T., Sallaberry, C., 2012. *Typage de noms toponymiques à des fins d'indexation géographique*, Revue Traitement Automatique des Langues, Vol. 53, Issue 2, pp. 143-176.

Han, J., Pei, J., Yin, Y., 2004. *Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach*, Data Mining and Knowledge Discovery, January 2004, Vol. 8, Issue 1, pp 53–87.

Jiang, D., Cai, F., Chen, H., 2018. *Location-Sensitive Personalized Query Auto-Completion*, in Proceeding of the 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Zhejiang, China, 25-26 August, pp. 15-19.

Lingling, D., Yuansheng, L., 2015. *Improvement and Research of FP-Growth Algorithm Based on Distributed Spark*, In Proceeding of International Conference on Cloud Computing and Big Data (CCBD'2015), Shanghai, China, 4-6 November, pp. 105-108.

Loustau, P., 2008. *Interprétation automatique d'itinéraires dans des récits de voyage*, PhD thesis, Université de Pau et des Pays de l'Adour.

Najadat, H. M., Al-Maolegi, M., Arkok, B., 2013. *An improved Apriori algorithm for association rules*, International Research Journal of Computer Science and Application, vol. 1, Issue 1, pp. 01-08.

Rocío, A.-M., Erick, L.-O., 2010. *Geo information extraction and processing from travel narratives*, Transforming the Nature of Communication, in Proceedings of 14th International Conference on Electronic, Helsinki Finland, pp. 363-373.

Sadikov, E., Madhavan, J., Wang, L., Halevy, A.Y., 2010. *Clustering query refinements by user intent*, In Proceedings of Proceedings of the 19th international conference on World wide web, WWW'10, pp. 841–850.

Sallaberry, C., Baziz, M., Lesbegueries, J., Gaio, M., 2007. *Une approche d'extraction et de recherche d'information spatiale dans les documents textuels – évaluation*, In Proceeding of Conférence en Recherche d'Information et Applications (CORIA), Saint-Etienne France, pp. 53-64.

Sanderson, M., Kohler, J., 2004. *Analysing Geographic Queries*, In Proceedings of SIGIR the Workshop on Geographic Information Retrieval, Sheffield UK, pp. 8-10.

Vargas, R.N.P., Moura, M.F., Speranza, E.A., Rodriguez, E., Rezende, S.O., 2012. *Discovering the Spatial coverage of the documents through the SpatialCIM Methodology*, AGILE'2012 International Conference on Geographic Information Science, Avignon, April 24-27, pp. 181-186.

Xueqing, L., Yangqiu, S., Shixia, L., Haixun, W., 2012. *Automatic taxonomy construction from keywords*, In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12), ACM, New York, NY, USA, pp. 1433-1441.

Yangqiu, S., Shixia, L., Xueqing, L., Haixun, W., 2015. *Automatic Taxonomy Construction from Keywords via Scalable Bayesian Rose Trees*, IEEE Transactions on Knowledge and Data Engineering, Vol. 27, Issue 1, pp. 1861-1874.