Geographical queries reformulation using a parallel association rules generator to build spatial taxonomies

Omar El Midaoui[1], Btihal El Ghali[2], Abderrahim El Qadi[3], Moulay Driss Rahmani[4]
[1,4]LRIT Associated Unit to the CNRST-URAC n°29, Faculty of Sciences, Mohammed V University in Rabat, Morocco
[1,2]SmartiLAB, Ecole Marocaine des Sciences de l'Ingénieur de Rabat, Morocco
[3]High School of Technology in Sale, LCS, Faculty of Science, Mohammed V University in Rabat, Morocco

Article Info

Article history:

Received Aug 15, 2020
Revised Oct 12, 2020
Accepted Jan 13, 2021

Keywords:

Big data
Geographical query
Information retrieval
Machine learning
Parallel FP-growth algorithm
Reformulation
Spatial entity

ABSTRACT

Geographical queries need a special process of reformulation by information retrieval systems (IRS) due to their specificities and hierarchical structure. 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 will be used in the reformulation of the spatial part of a geographical query. This approach exploits the documents that are in top of the retrieved list 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 frequent pattern growth (FP-growth) is applied to this database in his parallel version (parallel FP-growth: 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 using the taxonomy constructed based on our proposed approach improves the precision and the effectiveness of the IRS.

Corresponding Author:

Omar El Midaoui
LRIT Associated Unit to the CNRST-URAC n°29
Faculty of Sciences, Mohammed V University in Rabat
Rabat, Morocco
Email: omarelmidaoui@gmail.com

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 [1]. Web users searching for information that are spatially located often require information, that are geographically specific, such as geographic terms in Web pages and user queries or even user location [2]. 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. Moreover, the hierarchical structure that form the geographical context and the personalized relationships between geographical objects are also ignored by most of the search engines, which provides access to a very large number of heterogeneous and distributed resources of information. 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, instead of places that are close to that specified place and that represent the intent of the user.

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 [3, 4] that has elaborated this task using external resources named “gazetteers”. As defined in the literature, a gazetteer is a dictionary or geographic directory whose inputs are names of places. Each entry in this dictionary may be associated with information such as belonging to one or more administrative structures (town, region and country), Physical characteristic (mountain, river and road), with its statistical data and geometric representation expressed in a geographic referential.

Other works proposes the categorization of these spatial named entities after identification. Such as, Bouamor that exploits a document structure [5] extracted from the collaborative encyclopedia “Wikipedia”. 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 the end of the article. Buscaldi and Rosso has also proposed a technique for spatial named entities categorization using the thesaurus Geo-WordNet [6].

In the other hand, some approaches aim more particularly to the disambiguation of recognized places names [7]. The ambiguity can be understood as a word or a phrase that has many meanings [8]. In this case, two types of ambiguities are to treat [9]: 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 [10] which, first landmark names of places but also searches for these terms in ontological resources to identify related terms, potentially geographic. Domain-Specific taxonomies [11] are also playing an important role in many applications for improving search results [12] or helping with query reformulation [13, 14]. More recently, different external resources have been also created and used. Such as the geographic markup language (GML) data used in the geographical information retrieval model, proposed by Fang and Zhang [15], which simplifies the process of information acquisitions by the extraction and analyzes of the attribute features, spatial features and structure features of the exploited GML data. This data that is being generated by special services and stored in semi-structured documents using the geographic mark-up language, which is a coding specification for geographic information that has been implemented based on the extensible mark-up language (XML) and standardized (ISO 19136-2007). However, geographic taxonomies’ and ontologies’ hierarchical structure made their strength specially when considering the link of adjacency between places [16].

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. The thematic part of the query is improved using a query expansion approach. A specific type of query reformulation methods [17], which expand each search query with meaningful related concepts that can be captured from a manually or automatically constructed knowledge structure to enrich the query so it can represent its intent more clearly.

The expansion technique used for improving the thematic part of a geographical query in this work, is proposed based on the work of Nakade *et al.* [18] which proposes a semantic query expansion method that retrieve relevant tweets. This method uses a thesaurus (from thesaurus.com) to search for synonyms for original search topics and reformulate a query by adding synonyms and search topics using parenthesis and OR operators. According to an evaluation of this approach based on a corpus of 35000 tweets, the overall retrieval performance was improved.

In this paper, the proposed 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.

The main contributions done in this approach are the use of the parallel FP-growth algorithm in a geographical context by filtering the input document’s terms to the absolute spatial entities of the country for which the taxonomy is to be created. It includes also the step of classification of a submitted query (geo/non-geo) and the separation of geographical and thematic entities giving a geographical query to reformulate in order to reformulate the two entities in two different manners.

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 [19] 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. This spatial part contains a spatial relationship of adjacency and a noun of a city for which we are constructing the branch of the taxonomy. The proposed approach is based on the parallel FP-growth algorithm.

The geographical query model used considers two types of spatial 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 absolute spatial entity (ASE). While complex spatial entities as "near London" are labelled as an relative spatial entities (RSE).

2.1. The FP-growth algorithm

The FP-growth technique [20] 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 [21, 22] 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 of 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, while reordering the database tuples according to F-list, the non-frequent items are removed, based on the calculation of the value of support for every item (1) and considering the frequent items as the items for which the support value is greater then the minimum support threshold (minsup). Then the reordered transactions are inserted into the FP-tree. The input of the growth part of the algorithm is the constructed FP-tree. Considering a set of items A, N is the number of transactions in the database, f(A) the frequency of A in a database and P(A) the probability of occurrence of A in the same database, The support of A is calculated using the expression (1):

$$support(A) = P(A) = f(A) / N \qquad (1)$$

Then, the FP-growth algorithm 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. These 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. Finally, association rules to use in decision support are generated from the resulted FP-trees. A rule is valid when its confidence (2) exceeds or equals a fixed minimum confidence "minconf". Considering A and B as sets of items the confidence that A can imply B is measured as (2):

$$Confidence(A \rightarrow B) = P(A|B) = Support\ (A \cup B)/Support(A) \qquad (2)$$

2.2. The parallel FP-growth

The parallelized FP-growth work on distributed machines [23]. Its partitioning task 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:

- 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 the parallelizing step as our final result.

PFP distributes the growing FP-trees work based on the transactions' groups. thus, this approach is more scalable than a single-node implementation. PFP is implemented in the machine learning library (Mllib) on Spark and it takes three parameters: The minimum support threshold to identify frequent itemsets, 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 Figure 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 build the taxonomy, the fusion of all the generated FP-trees is forming our geographical taxonomy.

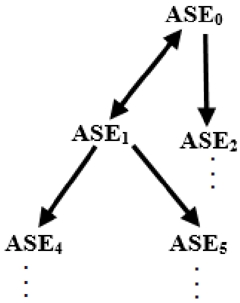

Figure 1. A two-level taxonomy for the ASE_0

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.

As shown by the example in Figure 1, ASE_1 involves ASE_0 ($ASE_1 \rightarrow ASE_0$) and ASE_0 involves ASE_1 ($ASE_0 \rightarrow ASE_1$). That means that these two absolute spatial entities are highly related. Thus we assume that they are close to each other geographically. In this case, the arc between the two ASEs is kept and this taxonomy evolves to a two-level taxonomy as it is the case in Figure 1. Otherwise, ASE_0 has involved ASE_3 ($ASE_0 \rightarrow ASE_3$), but ASE_3 do not involve ASE_0 ($ASE_3 \nrightarrow ASE_0$) so this arc has not been validated. Thus, it has been removed from the taxonomy. The double involvement is considered as a validation of the information generated by the descending implication.

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 [24]. 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. What remains of the query is marked as its thematic entity (TE).

We also proposed a contribution in this step. We made some modifications in the GIE approach cited 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 a relation of adjacency, we reformulate the spatial part of the query using the country's taxonomy [12], and the thematic part using a semantic expansion method. Our query reformulation approach, has been inspired from the work of Nakade *et al.* [18]. 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 it by the direct child-nodes items (CNIs) of the query's ASE in the geographical taxonomy as:

– User new query = TE SR ASE
– Reformulated query = expanded TE + "CNI_1" or "CNI_2" or …

In the reformulated query, 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 of 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 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 took 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 considered as a root for 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 selected list of 50 cities and villages of Morocco). For every ASE, we selected the thirty first web pages retrieved when submitting a relative spatial entity 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 of writing cities' names by the persons who wrote the documents contents. Because, we have noticed before that the miss-matching problem arise particularly in the case of nouns that contain accents [25]. 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:

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 database. After that, we varied the minimum support from 0.2 to 0.8 without the validation step, while we fixed the minimum confidence to 0.6 as shown in Table 2. Later we computed the error rate and the number of rules generated in every case.

Table 1. Spatial relations

Annotation	Expression	Translation of the expression
SR 1	à côté de	Next to/beside
SR 2	à la périphérie de	in the periphery of
SR 3	à proximité de	Close to/to the periphery of
SR 4	aux alentours de	Around
SR 5	aux environs de	Surroundings
SR 6	les environs de	the surroundings of
SR 7	près de	Near to

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

	Error rate				Number of generated rules			
RS\minsup	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 build the taxonomy of adjacency, based on a minimum support of 0.4 and a minimum confidence of 0.6.

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. ASE$_1$→ASE$_2$ and ASE$_2$→ASE$_1$). Table 3 shows that the result given by the case of the average support solves the problems mentioned for the SR 3 and SR 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

Spatial relation	Error rate			Number of correct rules		
	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 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. We also varied the value of minimum confidence, and noticed that the best results are given using the first fixed value 0.6. Using the favorable conditions represented we continue the construction of Morocco’s taxonomy with 0.4 as a minimum support, 0.6 as a minimum confidence and using the average of support for validating links as shown in Figure 3.

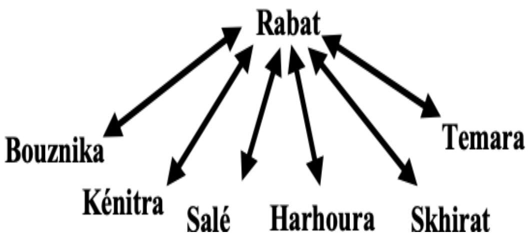

Figure 2. A one-level taxonomy for rabat using the spatial relation “Près de”

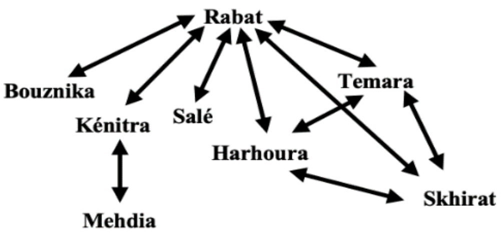

Figure 3. A two-level geographical taxonomy of adjacency for Morocco

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. We have been using multiple numbers of nodes also, in order to evaluate the effectiveness of our parallel algorithm. Thus, we took the baseline as the utilizations of one node with two cores of 2.3 GHz and 8 G of memory, this test has been conducted in 1.69 seconds. the second node’s characteristics are 2 cores of 2.3 GHz and 16 G of memory, the third node’s characteristics are 2 cores of 2.8 GHz and 4 G of memory, while the fourth node’s characteristics are 2 cores of 3.1 GHz and 16 G of memory. Table 4 shows the total execution time of our technique while varying the number of used nodes from 1 to 4, in consequence the number of cores, has been varied from 2 to 8 and the RAM capacity from 8 G to 44 G.

Table 4 shows that in term of execution time the effectiveness of our parallel technique increases while increasing the number of cores. Considering the baseline, we had a speedup of 3.13 when using 4 nodes. However, we notice in Figure 4 that this improvement is not proportional to the number of nodes used. Thus, we expect that at some level the efficiency of the algorithm will stagnate.

Table 4. Execution time (seconds) of the PFP algorithm on a spark cluster

# nodes	Baseline	2	3	4
# of cores	2	4	6	8
Memory (Go)	8	8+16	8+16+4	8+16+4+16
Duration	1.69	0.85	0.71	0.54

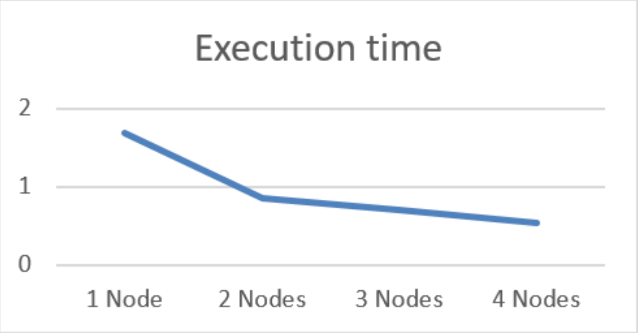

Figure 4. Execution time

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 built using the PFP algorithm and the same queries when reformulated using the geographical query expansion method (GQEM) proposed in [26]. Which is a natural language processing (NLP) method that modify and/or expand both the thematic and geospatial parts.

We compared the values of the Precision at 10, the Mean Average Precision and the execution time of the new proposed approach with the two cases: queries without reformulation and queries reformulated using GQEM. The performance is presented in percentage in order to show the enhancement of the precision and effectiveness of the proposed approach. The percentage of improvement of the execution time have been measured based on Amdahl's law.

From Table 5 we notice that the approach presented in this manuscript gives an interesting improvement in the precision of the geographical queries used in our experiments. However, the enhancement of performance between the new approach and GQEM given the execution time is quite normal, it is due to the parallelisation of the process. In order to evaluate our approach in other geographical places, we conduct tests on cities from France using the same parameter’s values. We took the capital Paris as a start. Applying our approach of adjacencies taxonomy builder using 120 documents in order to retrieve adjacent places to Paris, gives the results illustrated on Table 6. As shown in Table 6, the percentage of correctness of the results is of 100. This is due to the height precision and the double validation of our approach.

Table 5. The performance rate of the presented technique using 4 nodes according to the original queries and the queries reformulated with GQEM

Compared to	Precision at 10	MAP	Execution time
Original queries	10.56%	12.92%	43.24%
GQEM	5.30%	7.68%	68.05%

Table 6. Resulted child nodes for a one-level taxonomy of France

Child nodes	Distance with Paris	Validated (Yes/No)
Aubervilliers	6.57 km	Yes
Bobigny	8.85 km	Yes
Les-Mureaux	34.54 km	Yes
Ligné	**319 km**	**No**
Liré	**3539.32 km**	**No**
Livry-Gargan	15.13 km	Yes
Noisy-le-Grand	16.38 km	Yes
Pantin	6.47 km	Yes
Saint-Denis	8.22 km	Yes
Saint-Ouen	5.78 km	Yes
Villepinte	18.72 km	Yes

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 algorithm, 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 and association rules. Then we constructed the Moroccan taxonomy using a minimum support of 0.4 and a minimum confidence of 0.6 and the spatial relation “près de”, because these conditions gave the best results during our experiments. The proposed technique of reformulation has been tested on 50 queries, with a geographical intent and thematic entities from different fields. These queries had been reformulated based on the spatial taxonomy of adjacency. Finally, we compared the results retrieved by the Precision at 10, the mean average precision and the execution time. The results show that the reformulation based on our proposed approach and using a small number of reformulation terms has improved the value of the used indicators significantly. Considering the experimental results, we conclude that the presented method is an efficient work that permit to interpret and improve the results and effectiveness of queries containing a spatial entity of adjacency. The enhancement resulted from the proposed method is due to the use of a data mining approach based on an association rules technique, the processing of the geographical data in a personalized manner, considering the hierarchic structure of this data using taxonomies and the parallelisation of the process in order to minimise the execution time. 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 traces.

REFERENCES

[1] M. Sanderson, and J. Kohler, “Analyzing geographic queries,” *Proceedings of SIGIR the Workshop on Geographic Information Retrieval*, Sheffield UK, 2004, pp. 8-10.

[2] D. Jiang, F. Cai, H. Chen, “Location-sensitive personalized query auto-completion,” *Proc. of the 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Zhejiang, China, 2018, pp. 15-19.

[3] A.-M. Rocío and L.-O. Erick, “Geo information extraction and processing from travel narratives,” *Proc. of 14th Int. Conference on Electronic Transforming the Nature of Communication, Helsinki Finland*, 2010, pp. 363-373.

[4] P. Loustau, “Interprétation automatique d’itinéraires dans des récits de voyage,” PhD thesis, Université de Pau et des Pays de l’Adour, 2008.

[5] D. H. Bouamor, “Extraction des connaissances à partir du Web pour la recherche des images géoréférencées,” *Proceedings of Conférence en Recherche d’Information et Applications (CORIA)*, 2009, pp. 519-526.

[6] D. Buscaldi and P. Rosso, “Using GeoWordNet for Geographical Information Retrieval,” *Workshop of the Cross-Language Evaluation Forum for European Languages (CLEF)*, vol. 5706, pp. 863-866, 2008.

[7] D. Buscaldi, “Toponym ambiguity in geographical information retrieval,” *Proc. of the 32nd Int. ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR’09*, New York USA, 2009, pp. 847-847.

[8] R. N. P. Vargas, M. F. Moura, E. A. Speranza, E. Rodriguez, S. O. Rezende, “Discovering the spatial coverage of the documents through the SpatialCIM methodology,” *Proceedings of the International Conference on Geographic Information Science AGILE'2012*, Avignon, 2012, pp. 181-186.

[9] A. Einat, N. Har'el, R. Sivan, A. Soffer, “Web-a-where: Geotagging web content,” *Proc. of the 27th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK, 2004, pp. 273-280.

[10] M. Gaio, V.T. Nguyen, C. Sallaberry, “Typage de noms toponymiques à des fins d’indexation géographique,” *Revue Traitement Automatique des Langues*, vol. 53, no. 2, pp. 143-176, 2012.

[11] S. Yangqiu, L. Shixia, L. Xueqing, W. Haixun, “Automatic taxonomy construction from keywords via scalable Bayesian rose trees,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 1, pp. 1861-1874, 2015.

[12] L. Xueqing, et al., “Automatic taxonomy construction from keywords,” *Proc. of the 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'12)*, New York, NY, USA, 2012, pp. 1433-1441.

[13] E. Sadikov, J. Madhavan, L. Wang, A. Y. Halevy, “Clustering query refinements by user intent,” *Proceedings of the 19th International Conference on World Wide Web, WWW’10*, 2010, pp. 841-850.
[14] S. Aloteibi and M. Sanderson, “Analyzing geographic query reformulation: An exploratory study,” *Journal of the Association for Information Science and Technology*, vol. 65, no. 1, pp. 13-24, 2014.
[15] C. Fang and S. Zhang, “Geographic Information Retrieval Method for Geography Mark-Up Language Data,” *ISPRS International Journal of Geo-Information*, vol 7, no. 3, p. 89, 2018.
[16] M. Kokla and E. Guilbert, “A Review of Geospatial Semantic Information Modeling and Elicitation Approaches,” *ISPRS International Journal of Geo-Information*, vol. 9, no. 3, p. 149, 2020.
[17] M. A. Raza, R. Mokhtar, N. Ahmad, “A survey of statistical approaches for query expansion,” *Knowledge and Information Systems*, vol. 61, pp. 1-25, 2019.
[18] V. Nakade, A. Musaev, T. Atkison, "Preliminary research on thesaurus-based query expansion for Twitter data extraction," *Proceedings of the ACMSE 2018 Conference (ACMSE ’18), Association for Computing Machinery*, New York, NY, USA, 2018, pp. 1-4.
[19] H. Enghoff, “What is taxonomy?-An overview with myriapodological examples,” *Soil Organisms*, vol. 81, no. 3, pp. 441-451, 2009.
[20] J. Han, J. Pei, Y. Yin, “Mining frequent patterns without candidate generation: A frequent-pattern tree approach,” *Data Mining and Knowledge Discovery*, vol. 8, no. 1, pp. 53-87, 2004.
[21] M. Al-Maolegi and B. Arkok, “An improved apriori algorithm for association rules,” *International Journal on Natural Language Computing*, vol. 3, no. 1, pp. 21-29, 2014.
[22] H. M. Najadat, M. Al-Maolegi, B. Arkok, “An improved Apriori algorithm for association rules,” *International Research Journal of Computer Science and Application*, vol. 1, no. 1, pp. 01-08, 2013.
[23] L. Deng and Y. Lou, “Improvement and research of FP-growth algorithm based on distributed spark,” *Proceeding of International Conference on Cloud Computing and Big Data (CCBD’2015)*, Shanghai, 2015, pp. 105-108.
[24] C. Sallaberry, M. Baziz, J. Lesbegueries, M. Gaio, “Une approche d'extraction et de recherche d’information spatiale dans les documents textuels-évaluation,” *Proceeding of Conférence en Recherche d’Information et Applications (CORIA*), Saint-Etienne France, 2007, pp. 53-64.
[25] O. El Midaoui, B. El Ghali, A. El Qadi, M. D. Rahmani, “Geographical Query reformulation using a Geographical Taxonomy and WordNet,” *Procedia Computer Science*, vol. 127, pp. 489-498, 2018.
[26] Perea-Ortega, et al., “Applying NLP techniques for query reformulation to information retrieval with geographical references,” *Proceedings of the 2012 Pacific-Asia conference on Emerging Trends in Knowledge Discovery and Data Mining (PAKDD'12)*, Berlin, Heidelberg, 2012, pp. 57-69.

BIOGRAPHIES OF AUTHORS

El Midaoui Omar received the French high school diploma in Mathematics in 2005. Then, the Bachelor degree in Mathematics and Computer Science in 2009 from the Faculty of Science of Mohammed V University in Rabat (Morocco). Obtained his Master degree in applied computer science in 2011 from the same Faculty. Currently, he is preparing his Ph.D. in the Laboratory of Research in Informatics and Telecommunications (LRIT), and he works as a permanent teacher at the engineer School “EMSI” in Rabat. His research interests include: Data Mining, information retrieval, software engineering and Big Data.

El Ghali Btihal received the French high school diploma in Mathematics in 2006. Then, the Bachelor degree in Mathematics and Computer Science in 2009 from the Faculty of Science of Mohammed V University in Rabat (Morocco). Obtained her Master degree in applied Informatics and Telecommunication in 2011 from the same Faculty. Then her Ph.D in July 2016. Is currently, working as an assistant professor at an engineer School in Rabat. Her research interests include: Machine Mining, information retrieval, Big Data and Business Intelligence.

El Qadi Abderrahim is currently Professor in Computer Science Department, High School of Technology of Sale, Mohammed V University in Rabat, Morocco. He received his Ph.D. from the faculty of sciences, Mohammed V University in Rabat, in July 2002. In June 2010, he obtained his HDR from the same Faculty, in the subject: “Information Retrieval and Query optimization in Data warehouse”. His research interests include: data mining, text mining, web usage mining, information retrieval, query expansion, query recommendation, semantic web, data integration, SQL query Optimization and Big Data.

Moulay Driss Rahmani is Full Professor of Computer Science in Mohammed V University in Rabat (Morocco) where he was Head of Department of Computer Science until 2015. He has over 25 years of teaching experience (Concurrent programming in Java and Scala, Graphic User Interface, Compilation, XML Technology). His area of research interest includes Business Process Modelling, Urban Modelling, Wireless Sensor Network, Car-following modelling, Smart cities.