Characteristics of Geographic Information Needs

Andreas Henrich
University of Bamberg
Germany
andreas.henrich@wiai.uni-bamberg.de

Volker Lüdecke
University of Bamberg
Germany
volker.luedecke@wiai.uni-bamberg.de

ABSTRACT

A Geographic Information Retrieval (GIR) system for answering geographic queries has to cope with various information needs, which have a wide range of contexts and implicit requirements. A user, for example, who is looking for a place to spend his or her holidays certainly has a different understanding of distance than a user looking for a bar in the city he or she lives in.

To get a better understanding of geographic information needs and their implications for GIR systems, we analysed real world (geographic) queries with regard to different facets of geographic references in queries. The results of this analysis are presented in this paper, the aim of which was a classification of the geographic aspects of information needs. We present empirical results and line out possible classification criteria, which could be helpful in designing GIR systems that are able to consider different semantics of geographic references in queries.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation, Search process*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Experimentation, Human Factors

1. INTRODUCTION

Any information retrieval system should meet the information needs of its users rather than technically process a formulated query. A Geographic Information Retrieval (GIR) system for answering geographic queries should therefore be able to handle queries according to the information needs behind them. Especially the correct way of computing a spatial similarity or proximity between the geographic focus of a document and a geographic reference in a query greatly depends on the context of the query or the implications thereof. A user, for example, who is looking for a place to spend his or her holidays certainly has a different understanding of distance than a user looking for a bar in the city he or she lives in.

Knowledge of common real world geographic information needs and an analysis of their implications can be very useful in several ways: Firstly, the default behaviour of a GIR system (like the parameters or distance function in the spatial ranking [1]) can be adjusted according to the most frequent user needs; usually you have to define a default anyway, so this knowledge helps tuning a system. Secondly, this data may help creating and validating a classification scheme for geographic references, which can in turn be exploited in a way that a GIR system may automatically determine the best way of computing spatial similarity for a given query.

One way to learn about geographic information needs is to analyse existing query logs, ideally from a working and well-known GIR system publicly available. Since such a log was not available to us, we used logs of a text-based search engine and filtered it for queries with a geographic context. We are aware that a real GIR system would provide much better data, since many users know that standard text search engines do not really support geographic queries, and thus they do not enter geographic queries they might have in mind. The mechanisms and general thoughts remain the same, though, and are easily applicable to any query log available.

2. RELATED WORK

There has already been some research on how to determine whether a query is geographic or not and how to handle geographic references in queries and documents. [14] consider a query geographic, if it includes a place name, any other locator like a ZIP code, adjectives of place like *American* or terms descriptive of location like *state*, *county* or *city*, a geographic feature or a direction. [11] view a geographic query as a `<what, relation, where>` triple. They propose simple techniques to detect and disambiguate locations in queries, one of which deals with textual query strings that need to be parsed. They describe an algorithm for that, which uses ontologies and blind relevance feedback for determining relation and location of such a query. [8] describe a geo-query as a combination of constraints on content with geographic constraints and point out that a query may also be geographic, if a geographic reference is the answer to it.

[3] examine the use of natural language expressions in documents and focus on detecting positioning expressions,

which consist of a spatial relation and a landmark. Based on that they propose a geographic search strategy using query expansion.

[17] describes a way to detect location(s) associated with a (geographic) query by using blind relevance feedback, which is similar to [10], where the authors use implicit locations to improve retrieval performance of a GIR system. [4] computes the geographical scopes of web resources, where they deal with extracting and disambiguating location references.

[15] looks at the usage of location names in language, pointing out that geographic references made by users are often vague and that "it is worth defining real user needs when evaluating GIR". A formal model for metric details for the description of natural-language spatial relations (*overlaps*, *splitting*) is presented in [5]. [2] looks at the context of geospatial database semantics with regards to Human-GIS interaction and so suggests that the actual information need of a user is important for query processing. [7] deals with what-relation-where-like queries and how ontologies may support retrieval of documents relevant to geo-queries by query expansion, considering relations like *near*, *outside* or *north* and remark that there are different possible interpretations of relationships.

Research has also been done on classifying and analysing web queries, like [12], wherein a classification scheme for web queries is presented, which consists of the facets *Ambiguity*, *Authority Sensitivity*, *Temporal Sensitivity* and *Spatial Sensitivity*. Like us, they propose to automatically classify queries and invoke the right actions accordingly, but this classification is not specific to geographic queries. [6] present a formal framework for the formal definition of spatial entities, relations and queries, wherein spatial queries are classified into topological, set-theoretic, and metric queries, on the basis of the kind of relationships between the query object and entities in the search space involved. [9] classifies geographic queries into local or global queries according to their locality, using blind relevance feedback. [18] looks at the number of words and characters and the distribution of place-names in geographic queries and uses this knowledge for query rewriting, especially for the geographic references of queries. [14] parses a text search engine query log for geographic queries and surveys the length of these queries and the use of spatial relationships therein. The authors further classify the queries according to the query object. While all of this is related to our work, our focus lies on the geographic information need behind a query rather than the queries itself and especially on classification criteria of the intentions of geographic references in queries.

3. EXTRACTING INFORMATION NEEDS FROM GEOGRAPHIC QUERIES

3.1 AOL 500k collection

For our experiments we used the *AOL 500k Collection*, which was made public by AOL in 2006 and was intended for non-commercial research (see [13], as suggested in the copyright-file of the collection). It was withdrawn by AOL after a public discussion, though, since this data was in fact controversial, because it allowed to personalise some of the queries. We are aware of the ethical issues involved, but we consider it a valuable source of information for the purpose of this research, which handles the data absolutely anonymous.

The collection consists of real user queries to AOL search in the period of March 1st - May 31st 2006. It contains about 36 million lines of data provided by 650000 unique users, while a line of data is made up of an anonymised userid, a query string in lower case letters, a time stamp of the query, as well as the item rank of a result, if the user clicked on the link, in which case the domain name of the result is included in the data. Since each click on a result of the same query leads to a separate line of data and since there are several exact duplicates in the data, the actual number of unique queries is about 10-20 million, depending on what is considered a duplicate.

3.2 Detecting geographic queries

The first task is to automatically detect those of all queries that are geographic. In section 2 we outlined some approaches proposed in literature, which shows that this is not a trivial task. Therefore we take advantage of the fact that we do not have to make a decision on every query, as would be the case in a running live GIR system handling user queries. Since our experiments focus on the content of geographic queries (their concept part) – and not on quantifying the proportion of geographic queries of all web queries – we can narrow down the number of queries to be analysed to those that are unambiguously geographic. The sample size is large enough for that.

We used several patterns for regular expressions to automatically identify different sets of geographic queries. Each pattern consists of a preposition (none, *in*, *from*, *to*, *near*, *at*, *of*) and a "toponym part", which classifies the patterns in three groups:

The first group of patterns (*TopoState*)is the strictest and leads to queries that are almost certainly geographic. It makes use of a style of geographic reference that is characteristic for the US: a toponym followed by a state-name or the two-letter abbreviation of a state (state qualifier sq):

```
<concept> <preposition> <toponym t> <sq t is in>
```

We used the data provided by the WorldGazetteer to get about 26000 place names in the US and their corresponding states to find potential toponyms and to make sure that the toponym-state combination is valid. The queries we got are all very likely geographic, like "part time jobs near lansing illinois", "motels in santa monica ca" or "dachshunds for sale in greenville sc". Applying these patterns we got a total of 225421 geographic queries.

A second group of patterns (*StateOnly*)contains a concept-query, an optional preposition and a state-name or -abbreviation, leading to queries like "lakes in kentucky" or "used cars il". A manual review of these results showed that the proportion of genuine geographic queries is still very high, though false positives exist, e.g. "secure id" or "grow old with me". A total of 683397 queries were found that way (221677 match those of group *TopoState*, though).

Lastly, we created patterns to find all queries that contained a potential toponym without regards to the corresponding state and with varying prepositions. Browsing through the 1.5 million results, it becomes obvious that queries like "johanna lindsey" (Lindsey is a location in Ohio) and "sleep comfort" (comfort Texas) are hardly geographic. Using prepositions the results look more promising, but still contain a lot of false positives. We call this group *TopoOnly*.

Table 1: Number of results per pattern

Preposition	TopoState	StateOnly	TopoOnly
in	76064	77250	121968
from	700	1105	3231
to	1613	1771	7338
near	930	72	491
at	648	166	3890
of	9554	27944	68309
none or any	225421	683397	1513247

Though we are well aware that all numbers here have to be treated with caution, table 1 shows the number of results for each pattern and preposition. Most users seem to know that this search engine does not support "near"-like queries, which is a reason we did not evaluate other spatial expressions like *surrounding*, *between* etc. The absolute number of (good) results is large enough, though.

3.3 Experiments

We wanted to get an understanding of what people are actually looking for when they use geographic queries. That knowledge may be useful for parsing queries (known patterns or likely geographically interesting objects), for implementing ranking algorithms [1] or for designing geographic search engines (knowing geographic information needs). For this paper we understand a geographic query as a query that contains a query object (generalised to a concept) and a geographic reference (relation and location).

Since we wanted to analyse the concept part of the queries regardless of the concrete locations, we stripped all geographic references off the queries. We then manually classified several thousand queries (only one annotator). Based on that real world data we wanted to create a classification scheme for geographic queries, which is described in section 4. This process was explorative and iterative.

Web queries consist of only a few words, so it is often impossible to comprehend the actual information need behind a query, or it may be ambiguous. We tried to concentrate on queries whose intends are fairly obvious. Those that are not were left out; others were left out due to the sheer number of queries that is barely manageable manually. This has to be accounted for when looking at the numbers presented, which always contain the number of queries not classified.

The first aspect to be looked at is the content of the concept query itself. We can differentiate between an object and a predicate dimension here.

The object dimension is quite obvious from the terms used in the query (like "car", "house", "dogs") and describes what the query is about. We generalised terms to concepts, so that "house", "home", "apartments" etc. were classified as "habitation". Finally, we assigned each query one of 19 concepts. We intentionally did not use classification groups like in [14] or [16], so we could achieve a more fine grained result. Buying cars or animals can certainly be considered to belong to "Commerce and services", but both are so frequent that we found it more interesting to list them separately.

Looking at "homes for rent" and "homes for sale", we can see two different predicates for the same object. While the object dimension is mostly explicit in the query, the predicate dimension is often implicit. For queries like "homes in alabama" it is not possible to tell whether the user wants to buy, sell, rent homes; he might even think of something completely different. Nevertheless, the predicate dimension itself is important, though not as meaningful as the object dimension and much more difficult to capture.

The next thing we analysed is the reason for the geographic reference. We identified four different semantics: Firstly, you can make a geographic reference because you want to get there ("to there"). Secondly, you may want to do something near your location, like walk in the park or go to a bar ("there"). The focus here lies on the action on-site rather than getting to a certain location, although getting to a specific location first is a necessary requirement. Thirdly, you may want to know something about a location ("about there"). Finally, you may want to get something from a specific location ("from there"). Like with the predicate dimension, it is not always possible to make a clear decision here.

Finally, we looked at the "coverage" of the geographic reference and the semantic behind it. The ranking process in geographic search engines has to deal with areas, overlaps and degrees of matches (cf. [5]). If a user specifies an area for his query, like drawing a polygon in a map or writing "in texas", he is either looking for results that cover all of the specified area (maybe newspaper articles, pollution, weather) or for single items of interest within or near that area (homes for rent, bars), in which case the overlap is unimportant.

3.4 Results

The experiments resulted in 21 sets of queries (see table 1, three groups of seven prepositions each), not all of which contain a satisfying fraction of really geographic queries. Especially the patterns of groups *StateOnly* and *TopoOnly* without any preposition lead to a lot of non-geographic queries, while the queries found by any pattern of group *TopoState* were mostly geographic.

As can be seen in table 2, group *TopoOnly* contains a lot of names (39 of the top 100 queries[1] are forenames and only 27 are probably geographic queries), while group *StateOnly* contains a lot of place names (like San Antonio Texas; 52 of the top 100 queries are place names). Using "in" as preposition leads to better but also fewer results that may even exclude some information needs that are not typically expressed with that preposition.

Therefore we will concentrate on group *TopoState* results with any or none preposition in the patterns in this paper. We are aware that certain queries cannot be expressed in this form (like "cities in Texas" or "laws in Oklahoma"), but our analyses of all 21 sets indicate that this set is quite representative for all geographic queries.

The first question was what objects or concepts of objects the queries were about. Figure 1 shows the distribution of the concepts we identified. Four concepts are clearly dominant, especially in the top 100 queries: Habitation, Accommodation, Spare Time and Information. The example queries in table 3 give a good idea of those concepts. The large number of queries not classified is a result of many

[1]When referring to the top n queries we always mean the top n aggregations of query instances – geographic references stripped off –, so if we found 1000 queries for "hotel" – regardless of the location stated in the query – this would count as one result, while the frequency of each query is considered in every figure, of course.

Table 2: Most frequent concept parts of queries per pattern group with no or any preposition

TopoState		StateOnly		TopoOnly	
concept query	#	concept query	#	concept query	#
hotels	2115	university	3925	university	6366
city	2053	state	2646	city	4595
homes for sale	1062	web	2602	wells	3607
jobs	981	map	2477	deal or no	3492
real estate	947	craigslist	560	map	3436
map	848	dmv	516	washington	2871
weather	763	jobs	507	barnes and	2620
apartments	680	ocean city	470	state	2485
restaurants	598	branson	430	e	2429
motels	507	blue cross blue shield	368	google	2411
yellow pages	483	savannah	355	john	2329
town	460	real estate	322	craigslist	2248
holiday inn	440	orlando	304	whitney	2121
white pages	382	dest	293	the	1958
houses for sale	345	six flags over	290	david	1936
houses for rent	337	gatlinburg	275	chris	1805
hampton inn	336	craigs list	274	jessica	1735
homes for rent	282	republican party	273	james	1622
new homes	280	homes for sale	273	people	1612
remax	263	hannah	267	boston	1542

Table 3: Top 5 queries for the four most frequent concepts

Habitation	Accommodation	Spare Time	Information
homes for sale	hotels	restaurants	map
real estate	motels	things to do	weather
apartments	holiday inn	ymca	zip code
houses for sale	hampton inn	movie theaters	newspaper
houses for rent	comfort inn	campgrounds	population

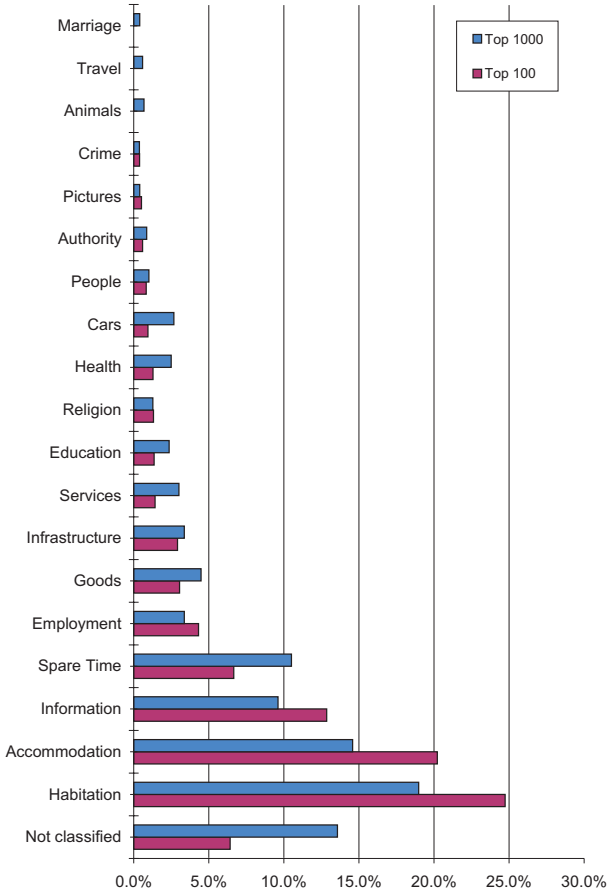

Figure 1: Objects of geographic queries of group *TopoState*

queries being ambiguous, like queries for directories (yellow pages, white pages etc.) or city names ("city of san antonio texas"), which are probably used in a second step to perform the actual query, thus the information need is not discernable. The distribution of frequencies in the top 100 and top 1000 queries are obviously different, which is partially because general objects like "hotel" are more frequent than each of all queries asking for car dealers of a certain brand ("dodge dealers", "bmw dealers" etc.).

We introduced the idea of a predicate dimension of a query. In figure 2 can be seen that buying or renting seems to be the dominant intention of users who make a geographic query. This is of course mainly related to the concepts Habitation and Accommodation, as well as Goods and Services. Since going there is a requirement for renting or buying things in most cases (not considering shipping of goods), a total of about 65% of the queries may imply that the user wants to actually get to the target location physically.

Another experiment looked at the intention of a geographic reference in detail. The frequency of the four intentions is shown in figure 3. As we assumed from the predicate di-

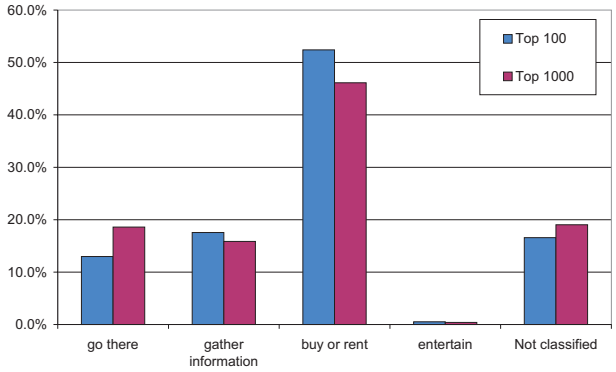

Figure 2: Predicates of geographic queries of group *TopoState*

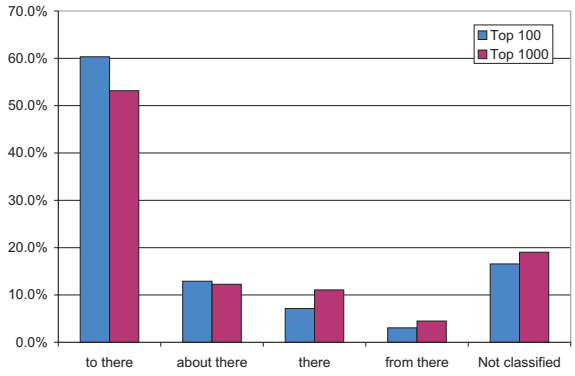

Figure 3: Intentions of geographic references of group *TopoState*

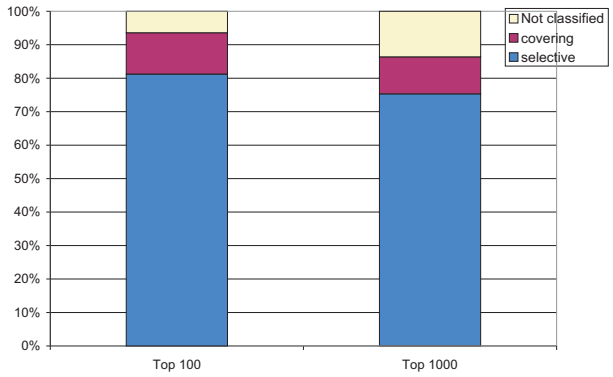

Figure 4: Intended coverage of geographic references of group *TopoState*

mensions, "to there" is by far the dominant intention. A geographic search engine should therefore probably be able to provide additional routing or travelling information. Only about 12% of the geographic queries are of pure informational character without the intent of going there.

Figure 4 shows that only about 12% of the queries are looking for documents that cover a whole area rather than single point(s). We considered queries for authorities (responsible for an area) covering, although they might as well be selective. It is not possible to make a clear decision here without actually knowing what the user intends. If he wanted to get to a concrete building to do some business there, a query for "department of children and family" would be selective. If he was looking for general information about the authority, the query would probably be covering. Queries for laws in certain areas are examples of pure covering queries.

These results are a clear indication that most of the (common) users' queries are selective, so that ranking algorithms of GIR search engines should take that into account. While complex representations of shapes of geographic foci and likewise complex algorithms for determining degrees of overlaps may be relevant to special queries, most common queries refer to single points.

4. ASPECTS OF GEOGRAPHIC INFORMATION NEEDS

Geographic information needs consist of at least a conceptual and a geographical component. Further characterising features can be identified in both components, so that it is possible to create a classification scheme, which could be used for a feature in geographic search engines: guessing the intention of a query and adapting the ranking process accordingly.

The conceptual part of a (geographic) information need consists of an object and a predicate dimension, which describe what the user is actually looking for (object) and what he intends to do with it (predicate). While there is a large number of possible values for the object dimension, which is usually expressed explicitly, there are only a few meaningful values for the predicate dimension, which is furthermore often hard to guess right. It is certainly neccessary to allocate single values to (aggregate) classes or concepts to manage the diversity of possible values.

Determining object and predicate of a geographic information need probably has a low direct impact on geographic search engines, but as we have seen in the preceding section, there is presumably a correlation between the concept of a query and its geographic aspects. Based on empirical data, it should therefore be possible to deduce information about the context of a query and the intention of a geographic reference and its implications.

We identified four defining features for the geographic aspect of an information need, two of which were introduced in section 3.

1. The *intention* of a geographic reference describes the relation between the user and the specified place, i.e. the user may want to go there, may want something from there, may be there or may want to know something about there.

2. *Coverage* means that the user is either looking for documents that are within or near a specified area (selective) or documents, whose content covers a certain area (covering).

3. While coverage refers to the content of a document, the *shape* of a geographic information need applies to the location of a document and determines the implicit similarity measures. In a query, a geographic reference typically is a point or a region (a city name is considered to be a "point"). Most other references may be interpreted as a region (course of a river, line between two locations etc.). Five distinct values for shape are possible: *at point*, *near point*, *within region*, *near region* and *route*. *At point* and *within region* specify hard constraints, whereas *near point/region* stands for fuzzy geographic constraints that need a distance measure. *route*-shapes are special cases where the information need is simply a routing request.

4. *Distance* is very dependant on the actual context. [2] describes a scenario with various interpretations of "near". Means of transportation, existing traffic routes and the intention of the user determine useful distance measures. During our analysis of the queries, we were not able to make decisions about the distance dimension of the information needs, so there is probably

Table 4: Classification of geographic information needs

Criterion	Possible Values
Object	any, typically classified
Predicate	any, typically classified
Intention	*to there*, *from there*, *about there*, *there*
Coverage	*covering*, *selective*
Shape	*at point*, *near point*, *within region*, *near region*, *route*
Distance	any

no way to extract distance information automatically from text queries. Nonetheless it is an important aspect that needs to be considered by geographic search engines, because it may have a large impact on spatial similarity measures.

In addition to that, one can think of further "meta dimensions" or "context dimensions" of information needs. Some queries, for example, are best answered by a yellow page service, while others are simply routing requests where dedicated routing and mapping services are the best solution.

Table 4 summarises our proposed classification scheme.

5. CONCLUSION AND FUTURE WORK

The results of our query analysis show that four concepts are clearly dominant in geographic information needs of common search engine users from the United States: Accommodation, Habitation, Information and Spare Time. Most queries implied that the user wanted to buy or rent something, so geographic queries also have a high commercial impact. The majority of users actually wanted to physically get to the location they were looking for, so routing and travelling information could be an important aspect of GIR search engines.

We described a possible classification or characterisation scheme for geographic queries, considering various intentions behind queries. While each aspect of this classification can be of importance, e.g. for (spatial) ranking algorithms for geographic search engines to compute degrees of coverage or distance measures, future work will be to derive geographic criteria of information needs from textual queries based on empirical data. Such a GIR system then could guess the right geographic ranking algorithms and better answer the users' queries.

6. REFERENCES

[1] L. Andrade and M. J. Silva. Relevance ranking for geographic ir. In *3rd Workshop on Geographic Information Retrieval, SIGIR'06*, Seattle, August 2006.

[2] G. Cai. Contextualization of geospatial database semantics for Human-GIS interaction. *GeoInformatica*, 11(2):217–237, June 2007.

[3] T. M. Delboni, K. A. V. Borges, and A. H. F. Laender. Geographic web search based on positioning expressions. In *GIR '05: Proc. of the Workshop on GIR, SIGIR 2005*.

[4] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *VLDB '00: Proc. of the 26th Intl. Conf. on VLDB*, pages 545–556, San Francisco, CA, USA, 2000. Morgan Kaufmann.

[5] M. J. Egenhofer and A. R. B. M. Shariff. Metric details for natural-language spatial relations. *ACM Trans. Inf. Syst.*, 16(4):295–321, 1998.

[6] L. D. Floriani, P. Marzano, and E. . Puppo. Spatial queries and data models. In I. C. A. U. Frank and U. Formentini, editors, *Information Theory: a Theoretical Basis for GIS*, pages 113–138. Springer-Verlag, LNCS, N.716, 1992.

[7] G. Fu, C. B. Jones, and A. I. Abdelmoty. Ontology-based spatial query expansion in information retrieval. In R. Meersman, Z. Tari, M.-S. Hacid, J. Mylopoulos, B. Pernici, Ö. Babaoglu, H.-A. Jacobsen, J. P. Loyall, M. Kifer, and S. Spaccapietra, editors, *OTM Conferences (2)*, volume 3761 of *LNCS*, pages 1466–1482. Springer, 2005.

[8] J. Graupmann and R. Schenkel. GeoSphereSearch: Context-aware geographic web search. In R. Purves and C. Jones, editors, *3rd Workshop on Geographic Information Retrieval, SIGIR'06*, pages 1–4, Seattle, 2006.

[9] L. Gravano, V. Hatzivassiloglou, and R. Lichtenstein. Categorizing web queries according to geographical locality. In *CIKM '03: Proc. of the twelfth intl. conf. on Information and knowledge management*, pages 325–333, New York, NY, USA, 2003. ACM Press.

[10] Z. Li, C. Wang, X. Xie, X. Wang, and W.-Y. Ma. Indexing implicit locations for geographical information retrieval. In *3rd Workshop on Geographic Information Retrieval, SIGIR'06*, Seattle, 2006.

[11] B. Martins, M. J. Silva, S. Freitas, and A. P. Afonso. Handling locations in search engine queries. In *3rd Workshop on Geographic Information Retrieval, SIGIR'06*, Seattle, 2006.

[12] B. V. Nguyen and M.-Y. Kan. Functional faceted web query analysis. In *Proc. of WWW 2007 Workshop on Query Log Analysis*, 2007.

[13] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *InfoScale '06: Proc. of the 1st intl. conf. on Scalable information systems*, page 1, New York, NY, USA, 2006. ACM Press.

[14] M. Sanderson and J. Kohler. Analyzing geographic queries. In *Proc. of the ACM SIGIR Workshop on GIR*, Sheffield, UK, 2004.

[15] D. Santos and M. S. Chaves. The place of place in geographical IR. In *3rd Workshop on Geographic Information Retrieval, SIGIR'06*, pages 5–8, Seattle, 2006.

[16] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic. From e-sex to e-commerce: Web search changes. *Computer*, 35(3):107–109, 2002.

[17] L. Wang, C. Wang, X. Xie, J. Forman, Y. Lu, W.-Y. Ma, and Y. Li. Detecting dominant locations from search queries. In *SIGIR '05: Proc. of the 28th annual intl. ACM SIGIR conf. on Research and development in information retrieval*, pages 424–431, New York, NY, USA, 2005. ACM Press.

[18] V. Zhang, B. Rey, E. Stipp, and R. Jones. Geomodification in query rewriting. In *3rd Workshop on Geographic Information Retrieval, SIGIR'06*, Seattle, 2006.