

# Challenges and Resources for Evaluating Geographical IR

Bruno Martins, Mário J. Silva and Marcirio Silveira Chaves  
Faculdade de Ciências da Universidade de Lisboa  
1749-016 Lisboa, Portugal  
{bmartins,mjs,mchaves}@xldb.di.fc.ul.pt

## ABSTRACT

This paper discusses evaluation of Geo-IR systems, arguing for a separate study of the different algorithmic components involved. It presents existing resources for evaluating the different components, together with a review on previous results in the area.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

## General Terms

Design, Measurement, Experimentation

## Keywords

Geo-IR, Evaluation

## 1. INTRODUCTION

There is an increasing interest over IR tools that access resources on the basis of geographic context, both in the academic and commercial domains (e.g. SPIRIT [19], mirago.co.uk or Google Local). However, the subject of Geo-IR is still at an early stage of development, and limited evaluation has so far been performed on such systems. Advances in the area require an evaluation methodology, in order to measure and compare different techniques [9, 22]. A Geo-IR track at CLEF2005 was established as an initial experiment – see <http://ir.shef.ac.uk/geoclef2005/>. However, a complete Geo-IR system involves different components, which interdependently influence one-another and could benefit from a separate evaluation. Some of the specific challenges are: 1) building geographical ontologies to assist Geo-IR; 2) handling geographical references in text; 3) assigning geographical scopes to the documents; 4) ranking documents according to geographical relevance; 5) building user interfaces for Geo-IR. Although an evaluation campaign like GeoCLEF can indeed be very helpful, there are many different variables at study, and the tasks involved are too complex to fit in such a general experiment. The objective of this paper is to highlight the feasibility of a scientifically sound approach for evaluating the different components of a Geo-IR system, whenever

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GIR'05, November 4, 2005, Bremen, Germany.

Copyright 2005 ACM 1-59593-165-1/05/0011 ...\$5.00.

possible building on standardized benchmarks, and separating the tasks associated with the challenges outlined above. We present existing resources for evaluating the different components, together with a review on previous experiments.

## 2. IR EVALUATION

Classic IR evaluation focuses on the main criterion of relevance. It uses measures derived from a contingency table dividing the data from a classification/retrieval problem into four distinct categories – see Table 1. Two very popular measures are precision and recall, which complement each other and usually incur in tradeoffs. For relevant items, recall  $r = \frac{tp}{tp+fn}$  is defined as the ratio of correct assignments by the total number of assignments. Precision  $p = \frac{tp}{tp+fp}$  is the ratio of correct assignments for relevant items by the total number of relevant assignments. Additional measures are accuracy  $a = \frac{tp+tn}{tp+fp+fn+tn}$  and error  $e = \frac{fp+fn}{tp+fp+fn+tn}$ , which are defined to be the ratio of correct and wrong assignments divided by the total number of system assignments. The  $f$ -measure combines recall with precision and is commonly used in problems where the negative examples outnumber the positive ones. The  $f1$ -measure equally weights precision and recall and is given by  $f1(p, r) = \frac{2pr}{p+r}$ .

	Relevant items	Irrelevant items
Labeled as relevant	true positives (tp)	false positives (fp)
Labeled as irrelevant	false negatives (fn)	true negatives (tn)

Table 1: Contingency table for binary classification problems.

Given two systems evaluated on the same test sets, we can determine whether one is better than the other using paired differences. This can be done through the Wilcoxon signed rank test, which uses ranks of differences to yield finer-grained distinctions than a simple sign-test [50]. The test imposes a minimal assumption, stating that the difference distribution is symmetric about zero (although empirical evidence suggests the test can be reliable even when this assumption is not met). In particular, the one-sided upper-tail test compares the zero-mean null hypothesis,  $H_0 : \theta = 0$ , against the hypothesis that the mean is greater than zero,  $H_1 : \theta > 0$ . To compute a statistic based on difference ranks, let  $z_i$  be the  $i^{\text{th}}$  difference, let  $r_i$  be the rank of  $|z_i|$ , and let  $\psi_i$  be an indicator for  $z_i$ , such that:

$$\psi = \begin{cases} 1 & \text{if } z_i \geq 0 \\ 0 & \text{if } z_i < 0 \end{cases}$$

The Wilcoxon signed rank statistic is  $T^+ = \sum_{i=1}^n z_i \psi_i$ . Upper tail probabilities for the null hypothesis are calculated for each possible value (i.e using the values from Table A.4 of Hollander and Wolfe [17]), and we reject  $H_0$  (and accept  $H_1$ ) if the probability mass is sufficiently small (0.05 is typically used as the threshold below which results are declared to be significant).

IR evaluation efforts such as TREC and CLEF use additional measures besides precision and recall. For ranked retrieval, it is common to compute precision and recall at fixed rank cutoffs (e.g. precision @ rank 10) or at fixed recall points (e.g. precision at 20% recall). Average precision is also commonly used, obtained by averaging the precision values at standard recall increases. Known-item searching (where users seek a specific resource) is evaluated through the reciprocal rank (RR) of the target document in the system’s ranked list. Thus, if a system retrieves the document at rank 4, the RR is 0.25 for that topic. If the target is not retrieved then the system receives 0 for that topic. A mean reciprocal rank (MRR) can be computed over all topics, and this measure has been reported as stable if a sufficiently large number of topics is considered.

Laboratory experiments may not represent real-world searching, and user satisfaction is often not highly correlated with traditional IR metrics [47, 51]. User studies can measure satisfaction with the output of the system, and this is particularly important when evaluating the quality of the user interface, a key element of successful retrieval system use. Although user studies require considerable effort to implement, previous works suggest that the percentage of usability problems detected in a test is approximately 31% per user, and that 80% of usability problems can be detected with only 5 users [33]. Typical IR user studies are task-oriented, having subjects using the system to find answers to particular problems. Although this has a strong quantitative component, it remains difficult to compare results between studies (each will have a different set of users) and between users (each subject will have his own standard of what constitutes a successful system). User studies also do not distinguish between ease of use and retrieval performance, and there are usually many variables involved.

Finally, computational aspects should also be considered for IR evaluation, since optimization is a key issue when handling terabyte collections. Ideally, algorithms should be no worse than linear, in order to effectively handle millions of documents. Hundreds of resources per second must be processed on a single workstation, which strongly affects the choice of heuristics to consider.

### 3. GAZETTEERS AND GEO-ONTOLOGIES

Gazetteers are important components of indirect geo-referencing through placenames [16]. Since Geo-IR usually relies on such external knowledge, algorithms depend on the gazetteer by one hand, and on the document collection by the other. The gazetteer is not simply an interchangeable component, as it gains reference status together with the test corpus [22]. Its data influences the outcome of any experiment, and it should therefore be carefully analyzed.

The limited availability of large gazetteers has been reported as a bottleneck [36], but freely available place lists are becoming more common, as many countries provide them in order to normalize the denomination of their cities – see [www.asu.edu/lib/hayden/govdocs/maps/geogname.htm](http://www.asu.edu/lib/hayden/govdocs/maps/geogname.htm) for a list of free place name gazetteers. For instance, UN-LOCODE, the official gazetteer by the United Nations, contains about 36000 locations in 234 countries. However, it is important to distinguish place lists (flat gazetteers) from geographic thesauri (here referred to as gazetteers or geo-ontologies), which besides listing places also provide hierarchic naming schemes and support at least some limited reasoning (i.e. topological or hierarchical relations). Flat gazetteers can suffice for tasks involving the recognition of geographical references, but other Geo-IR tasks require additional information. Spatial data in gazetteers is usually confined to centroids, which may seem too limited for determining spatial relationships. Nonetheless, we can use hierarchical/semantic relations, or combine them with Euclidean distances between centroids, to create hybrid reasoning meth-

Portuguese ontology		Multilingual global ontology	
Ontology component	Value	Ontology component	Value
Features	418,065	Features	12,293
Names	418,460	Names	14,305
Relationships	419,072	Relationships	12,258
Feature types	57	Feature types	7
Part-of rels.	418,340	Part-of rels.	12,245
Equivalence rels.	395	Equivalence rels.	1,814
Adjacency rels.	1,132	Adjacency rels.	13
NUT1	3	ISO-3166-1	239
NUT2	7	ISO-3166-2	3,979
NUT3	30	Agglomerations	751
Districts	18	Places	3,968
Islands	11	Admin. divisions	3,111
Municipalities	308	Capital cities	233
Civil Parishes	3,595	Continents	7
Zones	3,594	Other	4
Localities	44,386		
Street-like	146,422		
Postal codes	219,691		
Coordinate pairs	3,932	Coordinate pairs	0

**Table 2: Statistics for the geo-ontologies we developed.**

ods. The semantics of hierarchical relations are more difficult to grasp (i.e. administratively distant) but their use intuitively makes sense. There are also approaches that extend the spatial reasoning capabilities of gazetteers, through Voronoi polygons based on coordinates [1], or through spatial indexes based on uniform grids [38].

The Getty Thesaurus of Geographic Names (TGN) is one of the best known gazetteers and is used in many IR studies (i.e. it is one of the sources for the SPIRIT geo-ontology). Compiled from different sources, TGN contains about 1 million places around the globe, including both political entities (e.g. nations) and physical features (e.g. rivers). The focus of TGN records are places, each identified by a unique numeric ID. Linked to place records are names (common and historical, and names in different languages), the place’s parent in the hierarchy, other relationships (equivalent and associative), geographic coordinates, notes, data-sources, and place types (e.g. inhabited place, state capital). There may be multiple broader contexts, making the TGN polyhierarchical.

As part of our Geo-IR research, we developed two geographical ontologies, by consolidating data from several public sources. One considers global geographical information in multiple languages, while the other focuses on the Portuguese territory in more detail [7]. Encoded in OWL, the ontologies can be used for evaluating Geo-IR algorithms in different settings, namely on scenarios concerning either very large or narrow regions (each case has different ambiguity problems). Both ontologies are available as public sources. Table 2 shows some statistics. The considered information includes names for places and other geographical features (in 4 different languages – English, Portuguese, German and Spanish), adjectives of place, place types, relationships among features, demographics data, and geographic codes (e.g. postal codes). Digital maps were used to obtain additional centroids and to derive qualitative spatial relationships. In the future, we plan on adding information from other domains (i.e. companies or Internet domains with a known geographical context), as this can be useful in inferring geo-scopes for Web resources.

Existing gazetteers vary in many dimensions (e.g. scope, completeness, correctness, granularity, balance and richness [22]), and there is no standardization on the formats, contents, or service interfaces. As a consequence, data cannot be easily shared among these resources. Geo-ontologies are nonetheless seen as the promise of the Geospatial Semantic Web, and recent efforts have addressed interoperability through the use of Semantic Web standards [12].

## 4. FINDING GEO-REFERENCES IN TEXT

Recognizing place names in text is a crucial precondition for assigning documents with geographic scopes [10]. Although named entity recognition (NER) is a familiar task within Information Extraction, the problem here is more complex, as we must normalize the information in order to specifically describe or even uniquely identify place names. This involves disambiguating references with respect to their specific type (e.g. city, country) and grounding them with features at a geo-ontology. We can nonetheless build on previous NER efforts, particularly in what concerns evaluation. The specific problems of handling ambiguity and deriving meaning from place references have also been addressed in the past [20, 39], although in a lesser extent. No general scalable solution has so far been published, and no gold standard for evaluation is available. An interesting idea would be the establishment of a joint evaluation effort, similar to the work done in NER but focusing on the recognition and disambiguation of geographical references over text [9].

Mikheev et al. discussed the importance of gazetteers in finding geo-references, showing that a simple matching of the input texts to place lists performs reasonably well [32]. Nissim et al. experimented an off-the-shelf tagger for recognizing place names in Scottish historical documents [34]. They achieved similar performances to state-of-the-art NER results (an  $f_1$ -score of 94.25%), but a preliminary experiment in recognizing specific types (i.e. cities) yielded a drop in performance of about 20%. Comparing previous experiments in disambiguating place references – see Table 3 – can be a problem, as systems vary significantly in the disambiguation performed and on the evaluation resources. For instance, some systems only classify references according to their correct type, while others also ground references to coordinates or to a gazetteer.

System	Classify	Ground	Evaluation Results
InfoXtract [24]	✓	✓	93.8% accuracy
Informedia DVL [35]	✓	✓	75% accuracy
Web-a-Where [2]	✓	✓	63.1-81.7% accuracy
Smith and Mann [44]	✓		21.82-87.38% accuracy
Schilder et al. [40]	✓	✓	74 % $f_1$ -score
KIM system [26]	✓		88.1% $f_1$ -score
Nissim et al. [34]	✓		$f_1$ -score around 75%
Leidner et al. [23]	✓	✓	-
Metacarta [37]	✓	✓	-

**Table 3: Different systems handling geo-references in text.**

The annotations in corpora used for evaluating NER – see Table 4 – can be extended with a minimal effort, in order to associate place references to the corresponding entries at an ontology. Leidner already reported some ongoing work on this direction, discussing the importance of such sharable evaluation resources [22]. Building on resources already used in NER evaluation also enables a separate study of the recognition (which should be comparable to performances in traditional NER) and disambiguation steps. We are currently working on extending the annotations in the resources given at Table 4. Besides newswire texts, we are also annotating place references in a small corpus of Web pages. This allows evaluating HTML-specific heuristics, while also providing indications

Newswire Corpus	Words	Entities	Precision	Recall
Portuguese (HAREM)	89,241	1,276	86.63%	87.22%
English (CoNLL-2003)	301,418	10,645	96.59%	95.65%
German (CoNLL-2003)	310,318	6,579	83.19%	72.90%
Spanish (CoNLL-2002)	380,923	6,981	85.76%	79.43%
Dutch (CoNLL-2002)	333,582	4,461	78.54%	80.67%

**Table 4: Newswire corpora used in NER evaluations and the corresponding best recognition performances achieved.**

Corpus	Recognition			Grounding		
	Pre.	Rec.	$F_1$	Pre.	Rec.	$F_1$
Portuguese (HAREM)	89%	68%	77%	-	-	-
English (CoNLL-03)	85%	79%	81%	-	-	-
Spanish (CoNLL-02)	83%	76%	79%	-	-	-
Portuguese HTML	90%	76%	82%	89%	76%	81%
English HTML	91%	75%	82%	90%	73%	80%
German HTML	79%	72%	91%	77%	70%	73%
Spanish HTML	86%	75%	80%	83%	72%	77%

**Table 5: Our results in handling geo-references in text.**

on the problems associated with Web documents. Table 5 shows our initial results in handling place references over text, using a system currently under development [28].

The pre-processing tasks involved in handling place names should also be carefully evaluated. Language guessing (which we use to select recognition patterns) is for instance a well known problem, with previous approaches reporting accuracy around 90% [27]. Segmenting text into individual units has also received some attention. Grefenstette reported 95% accuracy using a high speed regexp method [14]. Mikheev reported error rates of 0.28% and 0.45% on the Brown and WSJ corpora, using rules like “when a period is preceded by an abbreviation and is followed by a lowercase word, proper name, comma or number, assign it as a sentence internal” [31].

## 5. ASSIGNING GEO-SCOPES

Besides recognizing geographical references in text, we can think in assigning documents to their corresponding geographical scope. This is a harder classification task, as multiple references (sometimes conflicting) can be associated with the same document.

Junyan et al. tried to classify pages according to three layers, namely nation, state and city. They used a hierarchical thesaurus, achieving an  $f_1$ -score of 86% [11]. Yamada et al. proposed to identify the geographical region mentioned in a Web page through a minimum bounding rectangle, reporting an accuracy of 88% [52]. However, these studies were evaluated over test collections specifically developed by the authors, making comparisons hard. A better alternative is the use of pre-existing human-made judgments, for instance Web pages in the Open Directory Project (ODP), after mapping the hierarchical organization of geographical classes in ODP to the geo-scopes used in the evaluated system. The `Top:Regional` branch of the directory is devoted to pages with a coherent geographical scope, containing about 1 million entries. Sub-branches of `Top:Regional` can be used to evaluate scope assignments at a high level of detail, as resources for some countries are categorized according to narrow regions. However, ODP pages cannot accurately model Web linkage, and this collection is therefore inappropriate to evaluate heuristics based on hypertext links. Amitay et al. proposed to find the geographical focus of Web pages when several place names are mentioned in the text, using the immediate parent in a hierarchically structured gazetteer [2]. ODP data was used for evaluation, and the correct continent, country, city or exact scope were guessed 96%, 93%, 32% and 38% of the times, respectively. Table 6 shows initial results for our scope assignment approach, which builds on an geo-ontology and a graph-ranking algorithm. The approach is described in a separate publication [29].

Other gold-standard sources are the well known Reuters 21578 and RCV1 collections. They contain news stories categorized into region codes, although these only correspond to broad regions (i.e. countries). Our scope assignment method achieved 92% accuracy over the 21578 collection, but we could not make comparisons since, to the best of our knowledge, there are no other published results on assigning geo-scopes to this data.

Multilingual global ontology : ODP Top:Regional		
Granularity Level	Measured Accuracy	
	Most Frequent Reference	Graph-Ranking
Continent	91%	92%
Country	76%	85%
Exact Scope Matches	67%	72%
Portuguese ontology : ODP Top:Regional:Europe:Portugal		
Granularity Level	Measured Accuracy	
	Most Frequent Reference	Graph-Ranking
NUT 1	84%	86%
NUT 2	58%	65%
NUT 3	44%	59%
Municipalities	28%	31%
Exact Scope Matches	34%	53%

**Table 6: Our results in assigning geo-scopes to Web pages.**

## 6. GEO-RELEVANCE RANKING

Finding resources concerning a specific region is very difficult with keyword matches. Geo-IR goes further than standard text search, handling `concept@location` queries (i.e. documents relevant with respect to some concept and region) and using, for instance, geo-scopes computed for each document. Relevance judgments now have two different dimensions (thematic and geographical), raising the problems of defining geographical relevance, finding appropriate metrics for its computation, and evaluating them.

In geographic space, “everything is related to everything else, but near things are more related than distant things” [46]. We can hypothesize that the spatial relevance of a location with respect to a query region increases with decreasing Euclidean distance between them [41]. Extent of overlap can also be used to measure spatial relevance and, for instance, the greater the overlap between a region and a query region, the greater the assumed relevance [3, 15, 49]. Besides spatial distance, we can define notions of topological distance between locations. Examples include adjacency, connectivity or hierarchical containment. Hierarchical measures can, for instance, use the number of non-common parents between a pair of places within the hierarchies to which they belong [43]. In fact, the problem of measuring similarity in hierarchical semantic structures has been extensively studied [25]. Combinations of semantic and spatial methods can also be used to create hybrid metrics [13], which in turn can be further combined with thematic similarity to create an integrated measure for relevance ranking in Geo-IR [18]. In a separate publication, we describe possible schemes for document indexing and retrieval using geo-scopes [30].

In SPIRIT, relevance was evaluated both through user questionnaires and with a set of pages extracted from a terabyte Web collection, each judged according to spatial and thematic relevance [6]. This latter methodology is based on TREC and CLEF. It has a long tradition in IR, and strong advantages such as reproducible results. Considerable effort is however required in designing such experiments. Given a long set of queries and a standard collection, the ranked lists of results are submitted for assessment by human judges on whether the documents are relevant or not. The submissions are finally evaluated using measures derived from precision and recall. The generation of queries for testing Geo-IR raises specific issues, as queries should test the capabilities that are not available in standard systems (e.g. imprecise regions, ambiguous place names and spatial relations). It is nonetheless possible to build on previous efforts. For instance, CLEF2005 included a pilot track on Geo-IR, using newswire texts from existing CLEF collections. Our system participated in GeoCLEF2005, but official results are not yet available. We also plan on using the GeoCLEF datasets to evaluate different strategies for Geo-IR ranking [30].

## 7. GEO-IR PROTOTYPE EVALUATION

Evaluating Geo-IR should holistically consider performance and user interactions, as these aspects are often not correlated. It has been noted that “the user interface must be evaluated on the basis of how well it meets the user’s needs along several dimensions, including informativeness, user friendliness, and response time” [45].

Fast indexing methods have been proposed in the past. Besides speed, these methods are also typically evaluated in terms of index size and supported operations. If the geo-scopes represent spatial footprints, then methods like R-Trees can be used to speed up geometric operations [4, 21]. Techniques have also been proposed for attaching spatial indexes to a regular text index, in order to merge keyword searches with spatial queries [48]. In a separate publication, we survey Geo-IR indexing approaches [30].

As for interactions, a Geo-IR prototype should be presented to potential user groups, in order to collect information about interactions and system functionalities. Important issues to evaluate are:

- Importance of multi-modal interfaces that allow users to describe locations textually (place names and spatial relations) or through a map (i.e. select and zoom regions of interest).
- The granularity in accessing resources according to geographical concepts (i.e. cities or streets). Shanon has previously discussed how the granularity of the answers to “where questions” depends on the reference points of the speaker and listener (e.g. where is the empire state building: In New York, in the US, or on the 34th street and 3rd avenue?) [42].
- Importance of different (possibly fuzzy) relationships between geographical concepts for retrieval (i.e. north-of, adjacent-to, near-to)
- Advantages of different presentation schemes (i.e. ranked lists or results clustered according to proximity).

Borlung described a user-oriented approach for evaluating interactive IR using short task descriptions [5]. Relevance is assessed by the users with respect to their task, and interviews/questionnaires may also be used to study usability aspects. A good design of the questionnaire is crucial in collecting reliable results, and standard instruments for evaluating user interfaces have been proposed in the past [8]. Still, as we discussed on Section 2, results from these studies can be hard to compare, and many different variables are often involved in a particular experiment.

User studies in SPIRIT were based on 8 subjects and 4 scenarios, taking as inspiration the Questionnaire for User Interaction Satisfaction [8]. The deliverables from this project constitute good references for future Geo-IR user studies, and an important conclusion is that users need to experience their own scenarios in order to truly evaluate a prototype system. We conducted 2 usability studies during the initial design of our Geo-IR interface. Through the obtained information, we are currently in the process of adding more advanced functionalities to our prototype system.

## 8. CONCLUSIONS

This paper discussed several aspects in Geo-IR evaluation, showing the feasibility of a separate evaluation on the different tasks involved. We listed existing resources and results from previous experiments, also showing initial results from a system we are developing. Additional studies are currently underway, building on the notions presented here. We plan on using statistical significance tests to compare different approaches and parameters. Computational aspects will also be given considerable detail, since optimization is a key issue when handling large collections.

## 9. REFERENCES

- [1] H. Alani, C. Jones, and D. Tudhope. Voronoi-based region approximation for geographical information retrieval with gazetteers. *International Journal of Geographical Information Science*, 15(4), 2001.
- [2] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-Where: Geotagging Web content. In *Proceedings of SIGIR-04, the 27th conference on research and development in information retrieval*, pages 273–280. ACM Press, 2004.
- [3] K. Beard and V. Sharma. Multidimensional ranking in digital spatial libraries. *Special Issue of Meta-data - Journal of Digital Libraries*, 1(1):153–160, 1997.
- [4] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. The  $R^*$ -tree: An efficient and robust access method for points and rectangles. In *Proceedings of SIGMOD-90, the 1990 Conference on Management of Data*, 1990.
- [5] P. Borlund and P. Ingwersen. Measures of relative relevance and ranked half-life: Performance indicators for interactive IR. In *Proceeding of the SIGIR-98, the 21st Conference on Research and Development in Information Retrieval*, 1998.
- [6] B. Bucher, P. Clough, H. Joho, R. Purves, and A. K. Syed. Geographic IR systems: Requirements and evaluation. In *Proceedings of ICC-05, the 12th International Cartographic Conference*, 2005.
- [7] M. Chaves, M. Silva, and B. Martins. A geographic knowledge base for text processing. In *Proceedings of SBBD-05, the 20th Brazilian Symposium on Databases*, 2005.
- [8] J. P. Chin, V. A. Diehl, and K. Norman. Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of CHI-88, the 1988 Human Factors in Computing Systems Conference*, 1988.
- [9] P. Clough and M. Sanderson. A proposal for comparative evaluation of automatic annotation for geo-referenced documents. In *Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004*, 2004.
- [10] I. Densham and J. Reid. A Geo-coding service encompassing a geo-parsing tool and integrated digital gazetteer service. In *Proceedings of the HTL/NAACL Workshop on The Analysis of Geographic References*, 2003.
- [11] J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of Web resources. In *Proceedings of VLDB-00, the 26th conference on Very Large Data Bases*, pages 545–556. Morgan Kaufmann Publishers Inc., 2000.
- [12] M. J. Egenhofer. Toward the semantic geospatial web. In *Proceedings of GIS-02, the 10th symposium on Advances in geographic information systems*, 2002.
- [13] S. Göbel and P. Klein. Ranking mechanisms in meta-data information systems for geo-spatial data. In *Proceedings of EOGeo-2002, the 2002 Workshop on Earth Observation and Geo-Spatial Data*, 2002.
- [14] G. Grefenstette and P. Tapanainen. What is a word, what is a sentence? Problems of tokenization. In *Proceedings of COMPLEX-94, the 3rd Conference on Computational Lexicography*, 1994.
- [15] L. L. Hill. *Access to Geographic Concepts in Online Bibliographic Files: Effectiveness of Current Practices and the Potential of a Graphic Interface*. PhD thesis, University of Pittsburgh, 1990.
- [16] L. L. Hill. Core elements of digital gazetteers: Placenames, categories, and footprints. In *Proceedings of ECDL-00, the 4th European Conference on Research and Advanced Technology for Digital Libraries*, September 2000.
- [17] M. Hollander and D. A. Wolfe. *Nonparametric Statistical Methods*. John Wiley & Sons, 1999.
- [18] C. B. Jones, H. Alani, and D. Tudhope. Geographical information retrieval with ontologies of place. In *Proceedings of COSIT-2001, Spatial Information Theory Foundations of Geographic Information Science*, 2001.
- [19] C. B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. van Kreveld, and R. Weibel. Spatial information retrieval and geographical ontologies: An overview of the SPIRIT project. In *Proceedings of SIGIR-02, the 25th Conference on Research and Development in Information Retrieval*, 2002.
- [20] A. Kornai and B. Sundheim, editors. *Workshop on the Analysis of Geographic References*, 2003. (held in conjunction with NAACL-HLT 2003).
- [21] R. Lee, H. Takakura, and Y. Kambayashi. Visual query processing for GIS with Web Contents. In *Proceedings of the IFIP TC2/WG2.6 6th Working Conference on Visual Database Systems*, 2002.
- [22] J. L. Leidner. Towards a reference corpus for automatic toponym resolution evaluation. In *Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004*, 2004.
- [23] J. L. Leidner, G. Sinclair, and B. Webber. Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HTL/NAACL Workshop on The Analysis of Geographic References*, pages 31–38, Edmonton, Alberta, Canada, May 2003.
- [24] H. Li, K. R. Srihari, C. Niu, and W. Li. Location normalization for information extraction. In *Proceedings of COLING-02, the 19th Conference on Computational Linguistics*, 2002.
- [25] Y. Li, Z. A. Bandar, and D. McLean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4), 2003.
- [26] D. Manov, A. Kiryakov, B. Popov, K. Bontcheva, D. Maynard, and H. Cunningham. Experiments with geographic knowledge for information extraction. In *Proceedings of the HTL/NAACL Workshop on The Analysis of Geographic References*, 2003.
- [27] B. Martins and M. Silva. Language identification in Web pages. In *Proceedings of ACM-SAC-DE-05, the Document Engineering Track of the 20th ACM Symposium on Applied Computing*, 2005.
- [28] B. Martins and M. J. Silva. Geographical named entity recognition and disambiguation in Web pages, 2005. (To Appear).
- [29] B. Martins and M. J. Silva. A graph-based ranking algorithm for geo-referencing documents. In *Proceedings of ICDM-05, the 5th IEEE International Conference on Data Mining*, 2005.
- [30] B. Martins, M. J. Silva, and L. Andrade. Indexing and ranking in Geo-IR systems. In *Proceedings of the Workshop on Geographic Information Retrieval at CIKM 2005*, 2005.
- [31] A. Mikheev. Document centered approach to text normalization. In *Proceedings of SIGIR-00, the 23rd conference on Research and development in information retrieval*, 2000.
- [32] A. Mikheev, M. Moens, and C. Grover. Named entity recognition without gazetteers. In *Proceedings of EAACL-99, the 9th Conference of the European Chapter of the Association for Computational Linguistics*, 1999.
- [33] J. Nielsen and T. K. Landauer. A mathematical model of the finding of usability problems. In *Proceedings INTERCHI-93, the 1993 ACM Conference on Human Factors in Computing Systems*, 1993.
- [34] M. Nissim, C. Matheson, and J. Reid. Recognising geographical entities in scottish historical documents. In *Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004*, 2004.
- [35] A. M. Olligschlaeger and A. G. Hauptmann. Multimodal information systems and GIS: The informedia digital video library. In *Proceedings of the 1999 ESRI User Conference*, 1999.
- [36] G. Petasis, A. Cucchiarelli, P. Velardi, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos. Automatic adaptation of proper noun dictionaries through cooperation of machine learning and probabilistic methods. In *Proceedings of SIGIR-00, the 23rd conference on Research and development in information retrieval*, 2000.
- [37] E. Rauch, M. Bukatin, and K. Baker. A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HTL/NAACL Workshop on The Analysis of Geographic References*, 2003.
- [38] W.-F. Rieker. Automated Retrieval of Information in the Internet by Using Thesauri and Gazetteers as Knowledge Sources. *Journal of Universal Computer Science*, 8(6):581–590, 2002.
- [39] T. K. Sang, E. F., and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-Independent Named Entity Recognition. In W. Daelemans and M. Osborne, editors, *Proceedings of CoNLL-2003, the 7th Conference on Natural Language Learning*, pages 142–147. Edmonton, Canada, 2003.
- [40] F. Schilder, Y. Versley, and C. Habel. Extracting spatial information: grounding, classifying and linking spatial expressions. In *Proceedings of the Workshop on Geographic Information Retrieval at SIGIR 2004*, 2004.
- [41] C. Schlieder, T. Voge, and U. Visser. Qualitative spatial reasoning for information retrieval by gazetteers. In *Proceedings of COSIT-02, the 2001 Conference on Spatial Information Theory*, 2001.
- [42] B. Shanon. Where questions. In *Proceedings of ACL-79, the 17th Annual Meeting of the Association for Computational Linguistics*, 1979.
- [43] M. Sintichakis and P. Constantopoulos. A method for monolingual thesauri merging. In *Proceedings of SIGIR-97, the 20th conference on Research and development in information retrieval*, 1997.
- [44] D. A. Smith and G. S. Mann. Bootstrapping toponym classifiers. In *Proceedings of the HTL/NAACL Workshop on The Analysis of Geographic References*, 2003.
- [45] J. Tague and R. Schultz. Evaluation of the user interface in an information retrieval system: a model. *Information Processing and Management*, 25(4), 1989.
- [46] W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234–240, 1970.
- [47] A. H. Turpin and W. Hersh. Why batch and user evaluations do not give the same results. In *Proceedings of SIGIR-01, the 24th conference on Research and development in information retrieval*, 2001.
- [48] S. Vaid, C. B. Jones, H. Joho, and M. Sanderson. Spatio-textual indexing for geographical search on the web. In *Proceedings of SSTD-05, the 9th Symposium on Spatial and Temporal Databases*, 2005.
- [49] D. Walker, I. Newman, D. Medyckyj-Scott, and C. Ruggles. A system for identifying datasets for GIS users. *International Journal of Geographical Information Systems*, 6(6), 1992.
- [50] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.
- [51] R. Wilkinson and M. Wu. Evaluation experiments and experience from perspective of interactive information retrieval. In *Working notes of Empirical Evaluation of Adaptive Systems workshop at Adaptive Hypermedia Conference*, 2004.
- [52] N. Yamada, R. Lee, Y. Kambayashi, and H. Takakura. Classification of web pages with geographic scope and level of details for mobile cache management. In *Proceedings of W2GIS-02, the 2nd Workshop on Web and Wireless Geographical Information Systems*, 2002.