Evaluating Geographic Information Retrieval

András Kornai

MetaCarta Inc., 350 Massachusetts Avenue, Cambridge MA 02139, USA
kornai@metacarta.com
http://www.kornai.com

In Memoriam Erik Rauch

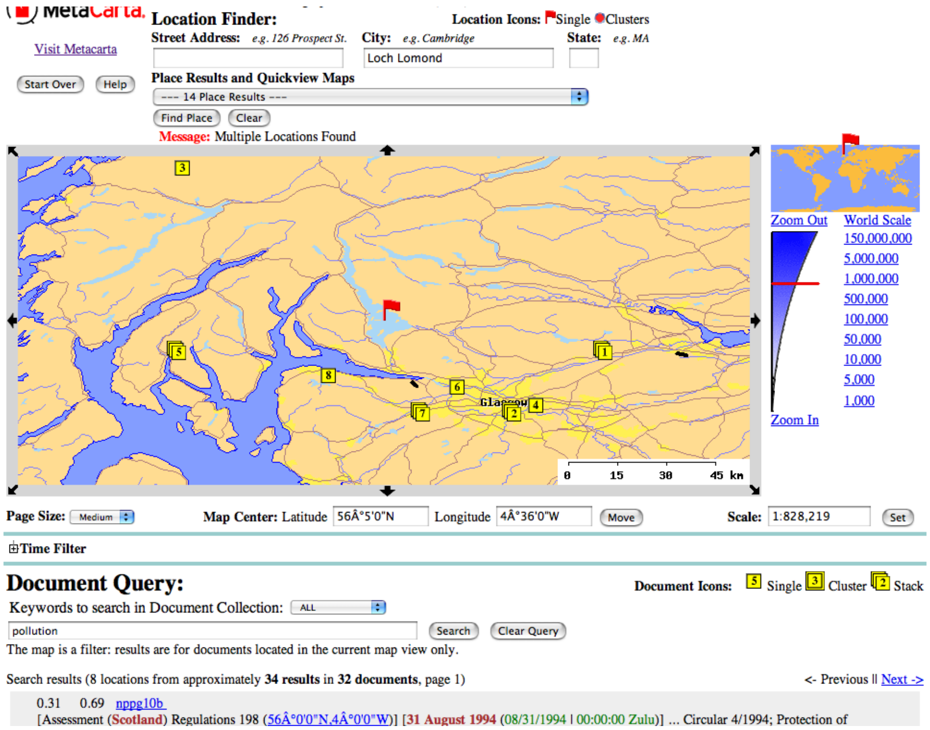
Abstract. The processing steps required for geographic information retrieval include many steps that are common to all forms of information retrieval, e.g. stopword filtering, stemming, vocabulary enrichment, understanding Booleans, and fluff removal. Only a few steps, in particular the detection of geographic entities and the assignment of bounding boxes to these, are specific to geographic IR. The paper presents the results of experiments designed to evaluate the geography-specificity of the GeoCLEF 2005 task, and suggests some methods to increase the sensitivity of the evaluation.

0 Introduction

The past 15 years have seen a great deal of controversy about the best way of evaluating Information Retrieval (IR) systems [Sherman:2000]. The *systematic* approach, developed in great depth at TREC [Harman:1993], is based on fixed collections, repeatable tasks, and uniform figures of merit, carefully keeping human judgment to an absolute minimum. The *user-centric* approach emphasizes the dynamic nature of the collections, the widely divergent paths that knowledge workers may take toward the same IR task, and the inherent difficulties in mapping user satisfaction to standardized figures of merit. This approach advocates detail tracking of individual "use cases" as the main avenue toward agile software development [Beck:2001]. While the cultural differences between the two groups are as large (and in many ways just as irreconcilable) as those between settled agriculturalists and hunter-gatherers, here we attempt the impossible and chart a middle course for the evaluation of geographic IR. Our starting point will be the MetaCarta user experience, which makes the map interface the focal point of the user's interaction with the data. Faced with a query such as the following:

> *Environmental concerns in and around the Scottish Trossachs.* A relevant document will describe environmental concerns (e.g. pollution, damage to the environment from tourism) in and around the area in Scotland known as the Trossachs. Strictly speaking, the Trossachs is the narrow wooded glen between Loch Katrine and Loch Achray, but the name is now used to describe a much larger area between Argyll and Perthshire, stretching north from the Campsies and west from Callander to the eastern shore of Loch Lomond.

the user selects a map region containing the Trossachs, and types in a few key phrases such as *pollution* or *environmental damage* or perhaps *tourism damage*. As document icons appear on the map, the user can rapidly scan the excerpts, recognize document stacks that contain documents referring to the exact same location, document clusters that refer to nearby events, and see isolated documents. There is no fixed discovery procedure: once the user gets an overall sense of the spatial density of pollution events in the region, she may decide to zero in on one subregion, perhaps one on the periphery of the original region of interest, perhaps one near the center.



Location Finder: Location Icons: Single Clusters

Street Address: *e.g. 126 Prospect St.* City: *e.g. Cambridge* Loch Lomond State: *e.g. MA*

Place Results and Quickview Maps: --- 14 Place Results ---

Message: Multiple Locations Found

Page Size: Medium Map Center: Latitude 56Â°5'0"N Longitude 4Â°36'0"W Scale: 1:828,219

⊞Time Filter

Document Query: Document Icons: 5 Single 3 Cluster 2 Stack

Keywords to search in Document Collection: ALL

pollution

The map is a filter: results are for documents located in the current map view only.

Search results (8 locations from approximately **34 results** in **32 documents**, page 1)

0.31 0.69 nppg10b

[Assessment (**Scotland**) Regulations 198 (56Â°0'0"N,4Â°0'0"W)] [**31 August 1994** (08/31/1994 | 00:00:00 Zulu)] ... Circular 4/1994; Protection of

Fig. 1. The MetaCarta interface

On the face of it, there is very little that is repeatable, let alone fully automated, in the discovery process: in particular, it would take very significant natural language parsing capabilities to derive two polygons that capture the "strict" and the "broader" Trossachs as defined above. In Section 1 we describe the processing steps we used, with special emphasis on whether we consider any given step relevant for geographic IR. In Section 2 we describe our experimental results, and consider the larger issue of whether the query texts require true geographical capabilities or are answerable by generic keyword search systems as well. In the concluding Section 3 we offer some suggestions how to make the evaluation task more specific to geographic IR.

1 Systematizing user-centric geographic IR

A single iteration of the MetaCarta geographic IR process consists of the user selecting a region (possibly the whole world) and a set of keywords (possibly empty), and viewing the results page. On this page, document icons are returned both on the map section and the textual section, the latter being similar to the results page of most major search engines. How the user proceeds to the next iteration seems very hard to model, especially as different users react to the same display with different strategies. *Geographic query refinement* is a subject of great intrinsic interest, and we will discuss some potential evaluation methods in Section 3, but here we confine ourselves to a single loop. Given a fixed collection of documents, such as the English dataset provided for GeoCLEF, a MetaCarta query has three parameters: `maxdocs` is the maximum number of document IDs we wish to see, typically 10 for "first page" results, **`bbleft bbright bbtop bbbottom`** are longitudes and latitudes for the bounding box, and an arbitrary number of keywords, implicitly ANDed together. To approximate a single iteration of the geographic IR process at least to some degree, we need to automatically set the `maxdocs` threshold (which we did by keeping it uniformly at 10), derive a bounding box, and select some keywords. Our first experiment was designed to assess the relative impact of the geographic versus the keyword component.

The queries can be submitted, with no geographic processing whatsoever, to a regular (non-geographic) IR system. In fact this was the strategy that the winning entry, the Cheshire II system [Larson:2005], followed. Since it was evident from the GeoCLEF topic set that the keyword component will have overwhelming impact, we decided that this factor is best controlled by standardizing on a less sophisticated, but widely available (open source) background IR algorithm: we chose Lucene [Hatcher:2004]. Further, we decided to standardize to a base level several preprocessing steps known to have significant impact on the outcome of IR evaluations. Since the goal was not to improve performance on the GeoCLEF task but rather to highlight differences between the geographic and non-geographic approach, the sophistication of these preprocessing steps was kept to an absolute minimum.

Defluffing We removed meta-guidance such as *find information about* or *relevant documents will describe* since the relevant documents will not have the words *relevant* or *document* in them. We call this step "defluffing" and perform it using a simple sed script that deletes the words *describing describe Provide provide articles article that discuss particular especially document relevant documents will describe Find Documents stories concerning give_detail statistics about report _any information items relating_to related_to especially _a _ing _s* by global search and replace. Notice that this step does *not* presume stemming or lowercasing, since we want to defluff irrespective of how we standardize these.

Stopword removal We defined stopwords as those words that had more than 1% of the frequency of the word *the* in a terabyte corpus we used for frequency analysis. This amounts to filtering out *0 1 A All For In S The This U What a about after all also and any are as at be because between but by can* *could e first for from has have if in including information is it its more much must name new not now of off on one or order other part people right should such take that the these they this time to two used was were where which will,* a total of 75 words.

Geographic defluffing We removed geographic metawords in a manner similar to defluffing: when the task description asks for countries involved in the fur trade the word *country* will not be in the docs. These words are *countries country regions region locations location Locations Location cities city.*

Stemming and lowercasing We performed neither stemming nor lowercasing, because the interaction of these operations with large sets of toponyms leads to many ambiguities not present in the original data. However, the possibility of using a standard (e.g. Porter) stemmer in conjunction with a large list of stemming exceptions (gazetteer entries) is worth keeping in mind. We are less sanguine about lowercasing, since the case distinction is a strong feature on proper names, and entity extraction without case information is noticeably harder.

Query expansion Vocabulary enrichment, in particular the local techniques pioneered by [Attar:1977] are now an essential part of IR. The geographic field also offers a particularly attractive way of expanding queries globally, since the hierarchical structure of geography, whereby *Oslo* is subordinated to *Norway* which is subordinated to *Scandinavia* which is subordinated to *Northern Europe* which is subordinated to *Europe*, is fixed once and for all. Here we performed neither local nor global query expansion, but we return to the matter in Section 3.

Query parsing While our overall strategy was to bring everything down to the lowest common denominator, and we performed no overall query parsing, we made a specific exception for Booleans, since these were often emphasized in the query text. For simplicity, we treated a query such as *Shark Attacks near Australia California* as two queries, *Shark Attacks near Australia* and *Shark Attacks near California* and merged the result sets.

After the steps described above, the topics (only `title` and `desc` fields kept) look as follows (autodetected geographic entities are shown in **boldface**):

001 Shark Attacks **Australia California** shark attacks humans
002 Vegetable Exporters **Europe** exporters fresh dried frozen vegetables
003 AI **Latin America** Amnesty International human rights **Latin America**
004 Actions against fur industry **Europe USA** protests violent acts against fur industry
005 Japanese Rice Imports reasons consequences first imported rice **Japan**
006 Oil Accidents Birds **Europe** damage injury birds caused accidental oil spills pollution
007 Trade Unions **Europe** differences role importance trade unions European
008 Milk Consumption **Europe** milk consumption European
009 Child Labor **Asia** child labor **Asia** proposals eliminate improve working conditions children
010 Flooding **Holland Germany** flood disasters **Holland Germany** 1995
011 Roman **UK Germany** Roman **UK Germany**

012 Cathedrals **Europe** particular cathedrals **Europe United Kingdom Russia**

013 Visits American president **Germany** visits President Clinton **Germany**

014 Environmentally hazardous Incidents **North Sea** environmental accidents hazards **North Sea**

015 Consequences genocide **Rwanda** genocide **Rwanda** impacts

016 Oil prospecting ecological problems **Siberia** and **Caspian Sea** Oil petroleum development related ecological problems **Siberia Caspian Sea**

017 American Troops **Sarajevo Bosnia Herzegovina** American troop deployment **Bosnia Herzegovina Sarajevo**

018 Walking holidays **Scotland** walking holidays **Scotland**

019 Golf tournaments **Europe** golf tournaments held European

020 Wind power Scottish Islands electrical power generation using wind power islands **Scotland**

021 Sea rescue **North Sea** rescues **North Sea**

022 Restored buildings Southern **Scotland** restoration historic buildings southern **Scotland**

023 Murders violence South-West **Scotland** violent acts murders South West part **Scotland**

024 Factors influencing tourist industry **Scottish Highlands** tourism industry Highlands **Scotland** factors affecting

025 Environmental concerns around Scottish **Trossachs** environmental issues concerns **Trossachs Scotland**

Table 1: Preprocessed Queries

Note how well the results of stopword removal from the `desc` section approximate the `title` section: aside from the last three topics, (where the `desc` section is really narrative) the two are practically identical. The stopword filtering step was included above very much with this goal in mind – in general, a good IDF weighting scheme will actually obviate the need for stopword filtering, but here we want to make sure that effects are not due to sophisticated integration of the different sections. This is not to say that such integration is worthless: to the contrary, its value is clearly proven by the Cheshire II experiments. However, we wished to take the narrative section out of consideration entirely, because the user-centric approach rarely, if ever, encounters queries of this sort, and we wished to make the results robust across the choice of `title` and `desc`. After these preprocessing steps, the queries are ready for submission to Lucene. Submission to the MetaCarta engine requires two further steps.

Identifying geographic references This task is generic to all geographic IR systems, and when we ran the 25 topics through the MetaCarta tagger we found that on the 124 geographic entities we had a precision of 100% (we had no false positives) and a recall of 96.8%: we missed *Scottish Islands* (twice), *Douglas*, and *Campeltown.* This suggests two evaluation paths: on the *discard* path missed entities are treated as plain (nongeographic) text, and on the *pretend* path we pretend the system actually found these. Either way (we found no significant difference between the two), the tagger is close enough to the ideal that the impact of geography is maximized.

Deriving bounding boxes Construed narrowly, this task may be specific to MetaCarta's query language: we use bounding boxes where others may use polygons, grids, tessellations, or other proximity schemes. Yet we do not wish to construe the task very broadly. In particular, we wish to exclude proximity schemes based on latent semantic indexing, hierarchical position in the gazetteer, or any other method that is entirely free of geographic coordinate information. MetaCarta computes bounding boxes offline (prior to having seen any query). For the experiments (including the submission) the following table was used:

Asia	25.0	179.9	6.0	
Australia	112.9	159.1	-9.1	-54.7
Bosnia Herzegovina	15.7	19.6	45.2	42.5
California	-124.4	-114.1	42.0	32.5
Caspian Sea	47.0	54.0	47.0	36.0
Europe	-11.0	60.0	72.00	32.00
Germany	5.8	15.0	55.0	47.2
Holland	3.3	7.2	53.5	50.7
Japan	122.9	153.9	45.5	24.0
Latin America	-118.0	-35.0	32.0	-55.0
North Sea	-4.0	8.0	65.0	51.0
Russia	26.0	60.0	72.0	41.1
Rwanda	28.8	30.8	-1.0	-2.8
Scotland	-8.0	0.0	61.0	55.0
Scottish Highlands	-8.0	-2.0	59.3	56.0
* Scottish Islands	-8.0	0.0	61.0	56.0
Siberia	60.0	179.9	82.0	48.0
* Trossachs	-4.5	-4.25	56.5	56.0
United Kingdom	-8.6	2.0	60.8	49.0
United States	-125.0	-66.0	49.0	26.0

Table 2: Bounding Boxes

Items marked by * did not have a bounding box in the database and reflect manual assignment, a fact that is reflected in our notion of **discard** versus **pretend** evaluation.

2 The main experiment

Though the point of the experiment is to compare pure keyword based IR, as exemplified by Lucene, to true geographic IR, as exemplified by MetaCarta, we did not think it appropriate to submit Lucene runs officially, and we submitted only the two pure MetaCarta runs of the five considered here. Needless to say, we used the same `trec_eval` settings to evaluate all five. In the following table, we summarize the `trec_eval` output for the five runs discussed in the text – for the definition of the various figures of merit run `trec_eval -h`.

Run #	**0**	**1**	**2**	**3**	**0+2**
Condition	MC geo	MC keywd	Luc bool	L w/o bool	Cmb MC+L
num_q	22	15	25	25	25
num_ret	1494	1002	820	500	1594
num_rel	895	765	1028	1028	1028
num_rel_ret	289	132	214	144	339
map	0.1700	0.1105	0.1819	0.1653	0.1959
R-prec	0.2155	0.1501	0.2328	0.2040	0.2396
bpref	0.1708	0.1148	0.1796	0.1570	0.1896
recip_rank	0.6748	0.6522	0.5453	0.5970	0.6778
ircl_prn.0.00	0.6837	0.6633	0.6064	0.6344	0.6878
ircl_prn.0.10	0.4178	0.2904	0.5096	0.4757	0.4505
ircl_prn.0.20	0.3443	0.2188	0.3748	0.3338	0.3740
ircl_prn.0.30	0.2977	0.1700	0.1622	0.1765	0.2986
ircl_prn.0.40	0.1928	0.1103	0.1161	0.1453	0.2064
ircl_prn.0.50	0.0971	0.0676	0.0976	0.1301	0.1221
ircl_prn.0.60	0.0435	0.0365	0.0687	0.0680	0.0750
ircl_prn.0.70	0.0261	0.0109	0.0687	0.0430	0.0597
ircl_prn.0.80	0.0130	0.0109	0.0663	0.0410	0.0457
ircl_prn.0.90	0.0000	0.0109	0.0513	0.0063	0.0207
ircl_prn.1.00	0.0000	0.0089	0.0394	0.0063	0.0194
P5	0.4455	0.3467	0.4240	0.4160	0.4640
P10	0.3182	0.2333	0.3680	0.3640	0.3560
P15	0.2667	0.1867	0.3627	0.3227	0.3067
P20	0.2500	0.1867	0.3300	0.2880	0.2820
P30	0.2182	0.1644	0.2360	0.1920	0.2427
P100	0.1141	0.0740	0.0856	0.0576	0.1204
P200	0.0636	0.0410	0.0428	0.0288	0.0660

Table 3: Comparing geographic to keyword search

For **Run 0** we only took the title words, the automatically detected regions, created a query as described above, with `maxdocs` set at 200. Since the system returns results in rank order, to create a first page one can just apply `head` to the result set. When the query implied logical OR rather than AND, we run the queries separately and sorted the results together by relevance. This way, run 0 mimicked a true geographic search where the geographic portion of the query is input through the map interface.

In **Run 1** we used MetaCarta as a pure keyword search engine, where everything, including geographic words, is treated just as a keyword (so the discard and the pretend paths coincide) and the bounding box is set to the whole world. As we expected, this is considerably worse than using geography (MAP 0.11 as opposed to 0.17 in run 0), but leaves some lingering questions.

First, experimenter bias: obviously MetaCarta has a vested interest in proving geographic IR to be better than pure keyword IR – in our eagerness to prove the point, have we perhaps dumbed down our keyword search techniques too much? Second, MetaCarta keyword search, much like Google, is designed to deal with very large document sets and short queries, and is therefore purely conjunctive: if a document does not contain all the keywords it doesn't even surface. To address both these issues, we decided to rerun the test with Lucene, an independent, disjunction-based keyword search engine.

Run 2 uses Lucene with default settings, but the additional benefit of Boolean resolution at query time: just as in Run 0, queries like *Roman cities in the UK and Germany* are broken up in advance as *Roman cities in the UK* and *Roman cities in Germany* and the result sets are merged. **Run 3** is the same, except for the benefit of this manual Boolean resolution: here the entire burden of query parsing is handled by the Lucene disjunction mechanism.

That some mechanism to handle disjunction is needed anyway for the GeoCLEF task, with its relatively small document set and relatively long queries, is evident from the fact that a purely conjunctive system such as MetaCarta did not return any results for a number of topics: obviously no shark attacks took place near both California and Australia, and no Roman city is both in Germany and England.

Run 0+2 is a simple attempt to remedy this defect, using MetaCarta results where available, and reverting to Lucene results for those queries where no MetaCarta results were returned. Remarkably, the use of geography boosts Lucene about as much as manual handling of Booleans did.

3 Conclusions

Overall, the 2005 GeoCLEF task was not one where geographic IR systems could really shine: the best results were obtained by pure keyword systems, and the top two systems, Berkeley [Larson:2005] and CSU San Marcos [Guillen:2005], both reported neutral and even negative effects from adding geographic information. By our own estimate, in systems tuned to this task, selectively disabling the classic (keyword-based) IR strategies as described in Section 1 leads to a factor of four greater loss in performance than selectively disabling the geographic component. Since this was rather predictable from reading through the topics, we felt a need to demonstrate that geography does help after all, and devised our experiment to prove this point, evident though it may be from the user-centric perspective, in the context of a systematic evaluation. From the experiment and the preprocessing leading up to it, several main components of geographic IR emerge that need to be more strongly exercised in future evaluations, we discuss these in turn.

First, the selection of geographic entities was limited, and most of them fit in what MetaCarta calls "Tier 1", a small set (2350 entries) of core place names whose approximate locations are known to everyone with a high school education. With the possible exception of the Scottish Islands (a class better defined by listing than by coherent geography) and the Trossachs (whose boundaries are clearly explained in the narrative task) a system with a small post-hoc gazetteer table could handle most of the questions: the only entries missing from the Tier 1 gazetteer are *Argyll, Ayr, Callander, Loch Achray, Loch Katrine, Loch Lomond,* *Perthshire, Scottish Islands* and *Trossachs*, and these do not even appear in the non-narrative sections.

Given that the problem of avoiding false positives is increasingly hard as we add more and more entities to the gazetteer, a task that encourages the use of trivial gazetteers will not serve the overall evaluation goals well. As it is, MetaCarta has an F-measure of 98.36%, which would be quite impressive, were it produced on a more realistic test set. Even within this limited set, one has the feeling (perhaps unsubstantiated, the guidelines didn't address the issue) that many of the toponyms are used metonymically. In particular, *Europe* seems to refer to the EU as a political entity rather than to the continent (see in particular topics 4 and 8).

It is not clear that a TREC-style evaluation like CLEF is the ideal forum for evaluating geographic coverage and disambiguation issues: clearly these can be measured more directly as part of a MUC-style named entity recognition task. One possible solution is to standardize on a single entity extraction tool; another is to distribute the extraction results as part of the train and test sets. Either way, it is important to realize that by taking large vocabulary issues off the table we artificially decrease the inherent difficulties of keyword techniques: with the multimillion word vocabularies typical of large gazetteers, the maintenance of good stemmers, obtaining reasonable background frequency estimates, and even correct tokenization are far bigger challenges than experience with small and medium vocabulary keyword-based IR would suggest. With large gazetteers, important multilingual issues such as phonetic spelling and exonyms crop up all the time, and it would fit the CLEF goals well to evaluate systems specifically in this regard.

Second, the issue of geographic proximity needs to be addressed in a more systematic fashion. In real life systems, a question about Hamburg may receive a relevant answer in a document that only discusses Bremen. We do not claim that the bounding box technique used by MetaCarta is ideal, and in fact we would very much like to see a task that would let us explore quantitatively the difference between alternative approaches. But it should be abundantly clear that tacking *in Rwanda* on a query does not make it truly geographic. The easy part of geography, continents and countries, is not any different from any other topic hierarchies. Continents expand to lists of countries rather trivially, but expanding Bordeaux to the list of over five thousand significant chateaux poses formidable knowledge engineering problems (and even if these are somehow surmounted, rare is the IR system that can handle a five thousand term disjunct over millions of documents gracefully).

This is not to say that the only real geographic queries are location questions like *Where was Osama bin Laden last seen?* – to the contrary, we find that even a small geographic hint as in *Bordeaux wine* or *Lexington preschool* is quite sufficient. Since such queries are in fact quite typical, parsing queries into geographic and non-geographic portions is an interesting research, and evaluation, topic. The 2005 descriptive queries offer a fascinating glimpse into problems that are viewed as important research topics such as negation *(Reports regarding canned* *vegetables, vegetable juices or otherwise processed vegetables are not relevant)*, or high level semantic reasoning (asking e.g. for *consequences, concerns, effects* and other highly abstract concepts generally considered beyond the ken of mainstream IR techniques). We do not deny the importance of these problems, but we question the wisdom of burdening GeoCLEF with these, especially as long as the simpler (but still very hard) query parsing questions remain unaddressed.

Finally, let us return to the question raised at the beginning of this article concerning the nature of the geographic query refinement loop. In the pure keyword search domain, the bulk of the work is spent on finding the right keywords: once these are at hand, at least in a well linked set of documents such as the web, both PageRank [Brin:1998] and hub/authority counts [Kleinberg:1999] provide sufficiently good results. In the geographic IR setting, typically there is no link structure (in this respect, the current document collection is very well chosen), and the only queries answered by purely geographic returns are the location questions. But the typical question is not about location, for the user knows it perfectly well at the outset that she is interested in wines from Bordeaux or preschools in Lexington. Rather, the bulk of the work is spent on analyzing the returns with some ordinal criteria (e.g. quality, price, trustworthiness, timeliness) in mind, and a realistic evaluation task would do well to choose a set of documents where some such criteria are easily computed.

Acknowledgements

We thank Doug Cutting for Lucene. Special thanks to Bradley Thompson, Karl Wright, Rachel Dillon, and the whole MetaCarta team, very much including Erik Rauch, whose sudden death fills us all with sorrow.

References

[Attar:1977] Attar, R. and Fraenkel, A.S.: Local Feedback in Full-Text Retrieval Systems. Journal of the ACM vol 24/3 pp 397–417

[Beck:2001] Beck, K. et al: Manifesto for Agile Software Development. `http://agilemanifesto.org`

[Brin:1998] Brin, S. and Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW7 / Computer Networks 30(1-7): 107-117 (1998)

[Guillen:2005] Guillen, R.: CSUSM experiments in GeoCLEF2005. This volume.

[Harman:1993] Harman, D.: Overview of the first Text Retrieval Conference In D. Harman (ed): Proc. 1st TREC, Publ NIST Gaithersburg MD. pp 1-20

[Hatcher:2004] Hatcher, E. and Gospodnetić, O.: Lucene in action. Manning Publications 2004

[Kleinberg:1999] Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM vol 46/5 pp 604–632

[Larson:2005] Larson, R.: Chesire II at GeoCLEF. This volume.

[Sherman:2000] Sherman, C.: Old Economy Info Retrieval Clashes with New Economy Web Upstarts at the Fifth Annual Search Engine Conference. `http://www.infotoday.com/newsbreaks/nb000424-2.htm`