A GEO-TEXTUAL SEARCH ENGINE APPROACH ASSISTING DISASTER RECOVERY, CRISIS MANAGEMENT AND EARLY WARNING SYSTEMS

Carsten Kropf, Shamim Ahmmed, Richard Göbel, Raik Niemann
Institute of Information Systems
University of Applied Sciences Hof
Alfons-Goppel-Platz 1 95028 Hof, Germany
shamim.ahmmed@hof-university.de, richard.goebel@iisys.de, carsten.kropf@iisys.de, raik.niemann@iisys.de
http://www.iisys.de

KEY WORDS: information retrieval, internet search engine, spatial search, toponym extraction, search rank

ABSTRACT[1]:
This paper presents an approach used in geo-textual search engines for application in security related domains like disaster recovery or early warning systems. Current approaches suffer from search conditions utilizing some combined scheme of textual and geographical search predicates. Standard retrieval engines support only either textual or spatial queries, like standard search engines or geographic information systems. One major problem regarding specialized search engine approaches is the extraction (geo-coding) and presen- tation of the related data. Especially, extraction of the desired information, such as geographical coordinates from unstructured texts and appropriate storage and retrieval techniques, is not covered by already existing web search engines. This paper presents a short introduction to a search engine architecture supporting a combined integrated approach of textual and spatial data derived from news websites of the internet and the retrieval of the data. Another point of interest of this paper is the presentation of the data acquired by a search query from the user. Specialized ranking algorithms take place in order to fit the results to the hybrid query combined by textual and spatial search predicates to figure out the best results to be presented. The search engine described here is capable of finding, storing and retrieving relevant documents posted by news websites to support, e.g., the work of disaster recovery teams in crisis areas. Besides the analysis of the given data this work also outlines the retrieval and ranking mechanisms which take influence on the relevance of a document towards the given search predicates.

1. INTRODUCTION

In the world of today, the Internet has become a very important source of information. Generally, whenever users begin to search something, they start by using internet web search engines1. These standard web search engines, however, are only capable of indexing textual data. These systems are commonly referred to as Information Retrieval (IR) systems.
One basic property of systems capable of assisting aid agencies or even military in crises areas is the relevance of places (points) of interest. Therefore, there are also information systems providing basic messages about certain areas. Yet, the basic problem still exists that these Geographic Information Systems (GIS) only support the storage, retrieval and appropriate presentation of geographical data like points or polygons.
During the last years, more and more research has been done to- wards information systems which can mix both types of data, tex- tual and geographical information. Such systems might search for up-to-date news postings on web sites which refer to some specific region of interest in order to support aid agencies. These up-to-date data might assist the helping forces by providing in- formation about current hot spots the forces would have to avoid or go to. Based on the combination of textual and geographical data, these systems are often called Geographic Information Retrieval (GIR) systems.
Hypothetically, they could assist teams in crisis management or early warning by searching for up-to-date information posted by news websites on the web, analyzing the given data and extracting helpful information about special regions from these data sources. However, gathering the desired information from news articles is not always easy, as there is a certain ambiguity inside the data. Mostly, no semantic markup is given inside the texts to be able to extract the spatial information the given article is related to. For this reason, special algorithms have to be applied capable of finding geographical points of interest inside the unstructured texts of the articles. The key challenge in this application domain is to find heuristics that enable the system to leach the given spatial references from the texts. These references, which are first present as place names (toponyms) in textual form, have to be assigned to an explicit coordinate on the world. This is why specialized algorithms have to be applied in order to identify these toponyms in a first step and assign the proper coordinate values in a second step, probably using a gazetteer. This assignment is called ”geo-coding”.
Typical problems for the geo-coding are place names (e.g. Bush) which have the same spelling as other words or place names which are not unique (e.g. Paris in France and Paris in Texas). As a consequence the geo-coding of unstructured texts usually produces imprecise results. This imprecision needs to be addressed by implementing one of the two following strategies:

- The geo-coding procedure may assign either the most likely candidate coordinates to a document. In this case every document refers only to a small number of coordinates.
- The geo-coding procedure may assign all but the most unlikely candidate coordinates to a document. Here every document will probably refer to a larger number of coordinates.

The second option was selected for the presented search engine to ensure, that almost all relevant documents are retrieved during a search (high recall). The disadvantage of this option is the large number of irrelevant documents contained in a search result (low precision). The presented search engine will compensate for this problem by an appropriate ranking function. This ranking function has to consider also geographical aspects providing the most relevant documents first. As an example documents with all geographical references

[1] This work has been funded by the Oberfrankenstiftung

contained in a search region are ranked higher than documents with a small percentage of geographical positions contained in this region.
Assigning all but the most unlikely coordinates to a document generates a second major challenge for the definition of an com- bined ranking function considering textual and geographical rel- evance criteria.
Therefore, a weighted sum of measurements calculated by these criteria has to be built to evaluate the quality of a news posting depending on the given search criteria. Based on the fact that the given system consists of multiple computers part of the challenge in this case is to present the most relevant articles collected from several computers.
The third key challenge in a GIR system is the management of the large datasets which occur in these systems. Specialized index structures and storage mechanisms have to be applied in order to keep a GIR system fast, especially regarding retrieval time. The storage mechanism has to be adopted to fit the given requirements of searching a set of input words and a combined search range in the geographical space. Therefore adopted storage and retrieval mechanisms have to be set up as well. These mechanisms are for example described in (Göbel, Henrich, Blank, & Niemann, 2009), (Göbel, Henrich, Blank, & Niemann, 2009) or (Göbel & Kropf, Towards Hybrid Index Structures for Multi-Media Search Criteria, 2010). This paper assumes the availability of such a mechanism and will not discuss them in further detail. In the following sections this paper describes the construction and algorithms of the Semi Automatic Research Archive Version 2 (SARA2) which is a special GIR systems supporting search requests for news combining textual and spatial search criteria. With these criteria SARA2 meets the requirements of a spatial search engine in crisis management supporting the retrieval of news in an area of interest. Furthermore the system can be easily adopted to display, e.g., hotspots in crisis areas. It can also serve as an early warning system that alerts users, when certain events occur by constantly analyzing incoming news.

2. RELATED WORK

The SARA2 search engine bases on results from a previous version, called SARA, developed in 2005 (Rill, 2005). However, the basic system was not planned to run on a multi processor cluster system. Therefore, only the basic algorithmic and logic could be taken as a basis for the second version of the architecture. Similar approaches were developed by the Spatially-Aware Information Retrieval on the Internet (SPIRIT) (School of Computer Science, 2011), (Jones, Abdelmoty, Finch, & Gaihua, 2004). However, this project was cancelled in 2005.
Basic geo-coding approaches are covered, e.g., by (Amitay, Har'El, Sivan, & Soffer, 2004). Other approaches to extract plain text from web sites and to provide basic toponym extraction from these plain text documents (which are already used in SARA2) are explained in (Helgert, 2009).
Text segmentation approaches such as stemming, explained in (Porter, 1980), or N-gram language detection (Dunning, 1994) are applied for the textual decomposition used as a preprocessing step for geo-coding and to store the full text in the database, properly.
Textual ranking bases on a cover density approach (Clarke, Cormack, & Tudhope, 2000).

3. SEARCH ENGINE ARCHITECTURE OVERVIEW

This section clarifies the basic architectural overview of the SARA2 search engine. The main parts of this architecture are summarized in Figure 1: Architectural Overview of the Search Engine.
This figure basically outlines three different layers in the architecture. As the entire software is supposed to run on a cluster system comprised by up to 50 nodes at the moment, there is a cluster base system that was also developed inside this project but is not part of the discussion, here. Basically, the SARA2 search engine software components work on the application level on top of the basic cluster management system. So, it can be assumed that there is some basic system managing, message handling and cluster node management.

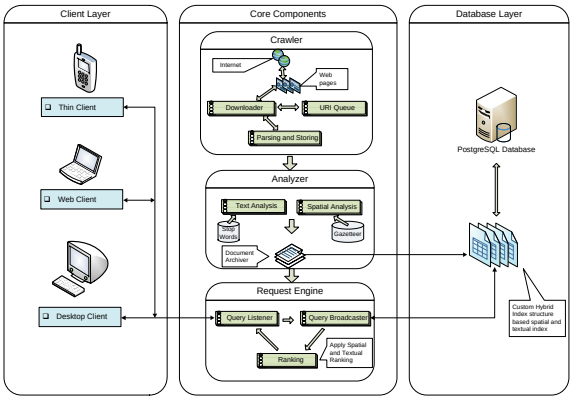

Figure 1: Architectural Overview of the Search Engine

Thus, this part only focusses on the "real" search engine component parts. Three different component parts are displayed in the figure:

1. Client Layer representing the client part of the search engine. These clients may use the engine only for retrieving information. Different kinds of clients, such as normal PCs as well as thin clients or mobile devices, may access the search engine via a predefined API.

2. Core Component Layer outlining the main search engine components including crawling, analyzing and retrieving documents. During the process of crawling, analyzing and storing articles, these components generate database entries and exchange information by the database.

3. Database Layer displaying the main storage area for the data. This layer contains a PostgreSQL database which is used to store the documents including the meta-data extracted during the crawl process. The database system has a specialized access method to the underlying data to ensure fast retrieval times on queries. However, this component is not focussed on in this paper.

This paper focuses on the core component layer providing the key functionality of the system. Basically, the crawl process described in item 2 combines crawling and analyzing. Therefore, some URI database exists which feeds the Crawler with URIs to crawl. These URIs are subsequently downloaded recursively and checked for their presence inside the database, which avoids downloading resources more than once. This check is supposed to be performed on the clear text extracted from the HTML sources, because of occurrences of dynamic elements inside the entire source code like datum values or commercials. If the entry for one resource is not flagged as being already present inside the system, this entry is marked for further processing. The Analyzer component fulfills the task of

analyzing the contents of the given resources downloaded before. For this reason, two separate analyses take place: the textual and the spatial one. The textual analysis, generally, tries to remove stop words from texts and to reduce the given words to their particular word stems. Spatial analysis uses some heuristics to extract toponyms from the unstructured texts and to assign these toponyms to unique coordinates using a Gazetteer which performs the conversion between place names and coordinates. The entire text processing procedure is explained in more detail in section 4. Having extracted the meta-information from the given documents they are subsequently stored inside a specialized document database using the Document Archiver component.
The Request Engine component provides searching access to the data stored inside the cluster system. Therefore, there is a component which broadcasts the queries into the cluster system and some component that waits for the responses from inside the cluster and collects the results. Some ranking algorithm using a combined scheme of texts and geographical data is applied to sort the entries by relevance. The details of the ranking process are described in section 5. This request component encapsulates the entire query process and the retrieval of entries in a predefined API which makes it easy to access it from different kinds of devices.

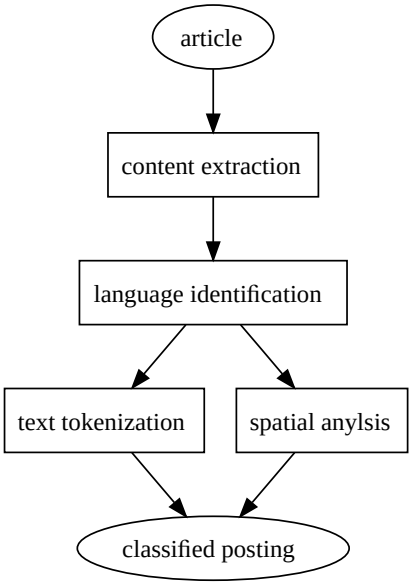

Figure 2: Internal Structure of the Text Analyzer Component

4. TEXTUAL ANALYSIS

As already discussed in the previous sections, the SARA2 geographical web search engine discovers news articles on web sites within a security related domain. Therefore, two different text analyses have to be performed in order to extract the plain text as well as the geographical coordinates of these articles. These articles come from web servers i.e. that the articles to analyze are only present in the semi-structured representation of HTML pages. For this reason, the algorithms applied here first extract the plain (main) text of the articles by applying heuristics inside the particular text passages and remove the markup given in the HTML pages. Standard approaches like tokenizing and stemming of the textual segments are performed on the full text part of the news articles. Subsequently a natural language processing approach is used to derive the toponym names from the full text and finally assign unambiguous geographical coordinates to the toponyms found inside the texts. “Geo-coding”, the procedure of extracting the toponyms and assigning explicit coordinates from texts is described in the following subsections. A general overview of the analyzing process is given in Figure 2: Internal Structure of the Text Analyzer Component.
It describes the process of textual analysis. First of all the clear text is extracted from the article and a language detection is applied. Thereafter, the spatial analysis of the textual content as well as the tokenization are performed. The information is then combined to a tagged posting document which is stored in the system.

4.1 TOPONYMS EXTRACTION

At present the search engine uses a toponym recognition algorithm for English. The applicability of this toponym extractor is checked by an algorithm which deduces the language of a document. This algorithm applies an N-gram based approach based on the distributions of sequences of n characters in different lan- guages (Dunning, 1994).

The algorithm for toponym extraction is described in (Helgert, 2009). Processing starts with the normalization of the given text. This activity includes the normalization of white space characters and the removal of these characters from the beginning or end of a text. Then, diacritical characters are eliminated and replaced by their base character (e.g. â → a). After the normalization of the text all place names are identified by using a gazetteer. Words from the text are only considered if they start with a capital letter. A place name may consist of more than one word. In this case the longest phrase from the gazetteer matching the current sequence of words is chosen as the place name. After the identification of these candidates for place names, several filters are applied stepwise removing improbable candidates. The filters can be separated in context free and context sensitive filters. So far the algorithm uses two context free filters:

- Remove all place names which are less than three characters long.
- Remove all place names which are part of a dictionary of some 100,000 common English words including stop words, colors, stems of common English words, etc.

After the context free filters, context sensitive filters are applied. These filters consider words before and after a candidate place name for their analysis:

- A place name needs to start with a capital letter everywhere in the text. If this is not the case, then the place name is removed.
- The last word of a place name consisting of multiple words needs to start with a capital letter. Exceptions are words which are introduced by a hyphen or an apostrophe.
- A hyphen or an apostrophe must not occur before or after a place name. The only exception is the ending “’s” for the last word of a place name.
- The place name is not to be included in quotation marks.
- Certain verbs like “lives“ and ”meets” must not occur before or after a place name. The only exception is the capital of a

country. This rule is controlled by two configurable lists of forbidden verbs before and after a term.

- No word starting with a capital letter is allowed after a place name. Exceptions are listed in a configurable list containing words like City, County or Valley. Note that a place name is removed from the list of candidates if only one of its occurrences in the text is followed by a word with a capital character.

- In general, no word starting with a capital letter may occur before a candidate place name. Exceptions are defined by a list of words like "at", "between" or "from". A second exception is a word occurring at the beginning of a sentence. In this case the word needs to be available in a dictionary of common English word (normalized via the Porter Stemming Algorithm) or the word occurs also in different parts of the text and starts with a small letter here.

- A place name consisting solely of capital letters may only start a new line or a new sentence. In this case the place has to be different from the initials of every phrase in the document.

- In front of a place name possessive pronouns like "my", "his" or "her" or articles like "a", "an" or "the" must not occur for every occurrence of the name in the document.

- The context of a place name needs to vary if the place name occurs at multiple places in the document. The context consists of two words in front of and two words after the document.

- In front of and behind the place name no phrase must occur which indicates that this name is a person. This rule is controlled by a configurable list of phrases like "father of" or "husband of".

A special rule applies if a place name is found at the beginning of a text. In this case, the place name is considered to be the unique place name for this document. The toponym extractor was tested with 500 randomly selected English texts taken from major news providers. Each text was manually checked and all relevant place names were marked. Afterwards the toponym extractor was applied to these texts and the results were compared. According to these tests, the extractor could identify 88.92% of all marked words (recall). Furthermore 88.23% of all words identified by the extractor were also manually marked (precision).

4.2 PLACE NAME DISAMBIGUATION

The toponyms extraction from the given articles only serves basic information about the place names which occur inside the texts. However, the basic toponyms extraction only produces place names, which means that there is no geographic reference on the coordinate systems the given place name is assigned to. One naive way to assign geographic coordinates to the given place names would be to look up the names inside a gazetteer. Yet, this assignment does not produce the desired results as it simply looks for candidate place names inside a given relation table.
It is obvious that the simple assignment of place names to coordinates is not the best way to allot a coordinate to a place name. There are a lot of ambiguous place names in the world, e.g. a place called "Paris" exists in France as well as in Texas. One important step in the textual analysis of the given news articles is, therefore, the unambiguous assignment of place names to coordinates.
For the disambiguation of the given place names, first, the toponyms extraction has to be completed successfully. Subsequently each toponym is passed to a gazetteer, taken from (Names, 2011), which then returns each coordinate asso- ciated with a given place name including duplicate values in different places. The "real" evaluation of the place name assignment is then checked as a weighted sum of different criteria. The list of criteria is dynamically constructed and can thus be adjusted to the current challenges of the application domain. After constructing the weighted sum of the distinct criteria the candidate entries are removed that underflow a given filter factor and thus do not satisfy, at least some of the criteria, sufficiently well to be taken into account as unambiguous coordinates for the currently inspected news article. The precision of the entire algorithmic may also be adjusted in order to provide flexibility. If each of the toponyms found inside the texts should be returned as geographical coordinate, this can be adjusted by applying a filter value.
The criteria applied to the toponyms and texts, respectively, are either based on certain events inside the given texts or simply based on a comparison of the particular toponyms and their posi- tions in the world. The criteria include:

- Country: checks whether a pre-defined country name is present inside the text of the news articles, besides the given toponym.

- First word: checks if the first word inside the full text of the article equals a certain toponym.

- Population: context free criterion which generally evaluates the population in a certain place name, represented by its associated toponym.

- Position: evaluates the position of a given toponym inside the news article, assigning a higher value if a toponym oc- curs in a front position of the text.

- Quantity: investigates the amount of occurrences of a given toponym in the particular article.

- Region: determines whether the region in which the given toponym is located in (based on the information of the gazetteer) also occurs inside the article.

- Subregion: assigns points for a given toponym if the subregion in which the toponym is located in also appears inside the text.

Each of the criteria mentioned above assigns a distinct value to the particular toponyms, including the distinct coordinates. In a second filtering step, each of the toponyms which underflow a certain threshold are removed preferring toponyms with a high resulting sum value of the particular distinct values resulting from the criteria.
Currently, there is no analysis given which inspects the spatial proximity or clustering behaviour of certain toponyms extracted from the given news article full text.

5. RANKING

Relevance ranking is a mission-critical part inside the search engine. Finding proper predicates and measurements for correct relevance ranking implies presenting more important results as first results returned from the search procedure. Furthermore, if

proper criteria can be found that apply ranking directly during querying the database, the work load may also be reduced for the entire system. Especially, sorting does not happen in the application layer, but in database layer, in terms of a stored procedure.
Relying on a scalable database system, this ranking procedure can also present the results without requiring too much main memory, which again reduces the minimum hardware requirements for the cluster system.
However, mainly criteria have to be found that represent the rel- ative quality of a document towards given search criteria. Applying the given combined scheme, results in separate score values for textual and geographical results, here. Approaches similar to PageRank, see (Page et al., 1999) which calculate the influence factor of a certain resource based on the references to or from this resource, are not applied here, because the number of links pointing to an article is not a sensible criteria in this application context.

5.1 TEXTUAL RANKING

The textual ranking is performed via the cover density approach (Clarke et al., 2000). This ranking approach measures the relevance of a given document based on the appearance of phrases which results in the fact that documents having a higher occurrence of each term will receive a higher ranking score. The occurrence measurement of individual terms is not taken into account as primary feature as the appearance of entire phrases is weighted higher than the occurrence of individual terms.
Two basic properties of the documents extracted in advance are passed into the particular ranking function and are weighted differently to produce a final ranking result:

1. The title of the news article

2. The tokenized full text from the news posting

The title in news postings (item 1) is, commonly, a very important property which relates directly to the content of the article. Generally, it can be assumed that the title in news posting directly relates to the topic covered in the article. Therefore, while searching for relevant documents inside the search engine, the title should be extracted in advance to be included in the rele- vance ranking. Obviously, the full text which is extracted dur- ing full text analysis of the news articles (item 2) should also be included in the search for specific keywords. Cover density al- gorithm is applied to score the candidate news postings using a scheme which applies weights to references in full text and the title of the web page. 20% of the total ranking influence are as- signed to occurrences of the given search keywords inside the title whereas occurrences inside the full text are weighted with 80% to form the final score value for the textual component of the rating. The weighting of the entries is calculated like this, because the cover density approach relies also on a ratio between the occurrences of phrases depending on the length of the texts. As the length of the title text, by trend, is low, the algorithm tends to apply quite high score values to the title, by default. Therefore, the title only gets assigned 20% of the total score value to limit the influence of the title towards the entire score value. Basically, these ranking criteria are grounded on the experiences made in conventional full text engines and information retrieval systems, such as web search engines.

5.2 SPATIAL RANK

Based on the fact that the relevance ranking has to be evaluated towards textual and geographical criteria, there exists also an approach to sort the particular candidate results by the spatial relationship of a posting according to the given search rectangle serving as input. The spatial ranking is grounded on ranking approaches from geographic information systems (GIS). However, these approaches cannot be transferred directly because in GIS in most of the cases, one particular point is related to one posting (or event). In the case of general geographical information retrieval systems, such as SARA2, there is a certain likelihood that more than one geographical coordinate is assigned to one news article posting. That is why, the relevance ranking criteria from GIS have to be extended to handle more than one reference.

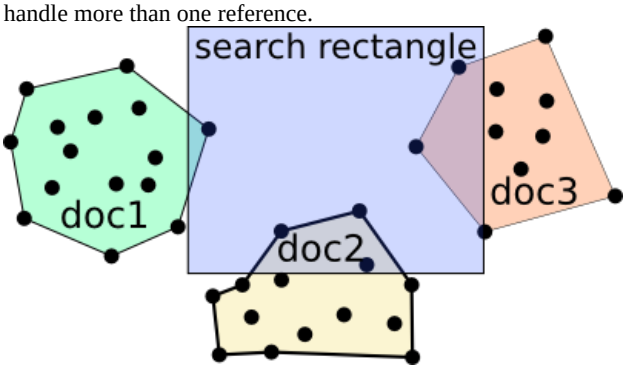

Figure 3: Search Rectangle Overlapping Three Document Footprints

This figure shows, generally, the document “footprints” of three articles as well as a rectangle used for searching related documents. The footprints represent a hull around the points contained in the particular documents. They are only displayed for keeping a better overview here and do not actually relate to the ranking or search procedure. While searching inside the entire set of documents each document containing points which reside somewhere inside the search rectangle are assumed to be candidates for fulfilling the search predicate. However, the fact that at least one point contained in a particular document has to be present inside the given search rectangle is the only measurement for one document to be a candidate. At this point, no quality evaluation is given, how relevant the document is regarding the search rectangle. Currently, the search results are evaluated based on the cover of the search rectangle towards the points contained inside the documents.
The spatial ranking approach evaluates the presence of a portion of the entire set of points contained inside one document posting in the given search rectangle. Therefore, it counts the amount of points that are inside the given search rectangle. Finally, the total spatial score is built by the following equation:

$$S = \frac{pts_{in}}{pts_{total}}$$

Equation 1
This equation calculates the total spatial score for one document as a ratio of the points that reside inside the search rectangle to the total points referenced from the document. More elaborate distance measures are also possible, which include, e.g., the distance from a particular center of gravity of the document footprint towards the center of the search rectangle or area overlapping evaluation between the convex hull of the document footprints and the search rectangle. However, even the basic ranking criterion, given in Equation 1

proved successful for the current state of the SARA2 spatial search engine.
Finally, the two independent ranking criteria have to be combined to present one final score value for a particular document which is then taken as input to the sorting function. The textual ranking is here taken with a weight of 60% and the spatial rank computed from the relation of the input search rectangle and the points contained inside the document is weighted with the remaining 40% because the textual ranking is supposed to have a higher priority, compared to the spatial impact, here.

5.3 RESULT COLLECTION

The ranking procedure takes place on each particular node attached to the cluster system. However, the search engine is supposed to display the best results first. These results are sent back from each of the particular cluster nodes. For this reason, they have to be combined from the entire set of documents returned from the distinct cluster nodes which only have knowledge about the subset of entries stored on the particular nodes. It is especially important that the ranking procedure on the nodes returns absolute values and does not rank the found documents relatively. Ranking the documents relatively would imply that they cannot be compared with documents originating from other nodes. If there is an absolute value assigned to the particular documents, the subsets from the distinct nodes may just be sorted according to their entire score value. So, the sorting procedure takes place on the node where the query is issued. In advance, this node has to broadcast the query throughout the cluster and prepare for receiving the results.
Besides the collection, there are certain other considerations made inside the cluster. For very common words, e.g. for news websites “news”, there is a very high likelihood that these words occur also in a large amount of documents on the cluster. However, based on the general setup as a web search engine, most people tend to view only a small amount of results before defining the search conditions more precisely as they understand that the given predicates are just too general to find the desired results. Thus, only small portions of the entire result set should be fetched at a time to reduce the total work load for the cluster system. Therefore, returning the search results is performed using a buffered approach. Each node returns the top 50 references not yet fetched at a time continuing to return the entries over time after the first top 50 are returned in the background. Using this approach, the entire set of documents may be transferred and waited for at once, however, there is also the possibility to show only the top results and proceed searching when a user request commands to do so.
A heuristic which is used to depict a rough estimation about the entire size of the result set exists as well. It generally issues the query planner to extract the estimated number of total results on one particular node.

6. CONCLUSIONS

In this paper, we described a search engine architecture which is usable for security related application domains. This search engine may have special impact on disaster recovery and crisis management as it has the ability of searching news articles from websites and extracting spatial information which is subsequently connected to the particular articles. Thus, this system can be used to identify hot spots in crisis areas where geographical data are especially important in connection with special events published in news articles. It can also be set up as an early warning system to signalize potential events based on heuristics and cluster based algorithm approaches.
However, the ranking algorithms, although they work properly, are still issued as a sorting algorithm outside the database storage and access mechanisms. Therefore, one future challenge is to include the relevance ordering inside the specialized database structures in order to reduce the work load of the application. Currently, work is done to improve the spatial assignment and text analyzing features which include, on the one hand, the extraction of the content text from unstructured articles as well as geo-coding of toponyms inside the documents to improve the accuracy of the textual analysis feature.
The application of the SARA2 search engine works properly and might be helpful in many application domains. The approach is flexible and can be used to solve many issues given in disaster recovery or early warning systems by applying the combined scheme of textual and spatial searches. However, the analysis has to be improved, still, to provide more stable algorithms which assign the place name, e.g., based on a spatial density. Some clustering algorithm would make the results more robust. Until now, only a small portion of users is able to use the search engine. So, another focus is, currently, to spread this application to broader public.

REFERENCES

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: geotagging web content. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (S. 273 - 280). ACM.
Clarke, C. L., Cormack, G. V., & Tudhope, E. A. (2000). Relevance ranking for one to three term queries. *Inf. Process. Manage.* (36), 291 - 311.
Dunning, T. (1994). *Statistical Identification of Language.* New Mexico State University.
Göbel, R., & Kropf, C. (2010). Towards Hybrid Index Structures for Multi-Media Search Criteria. *DMS*, (S. 143 - 148).
Göbel, R., Henrich, A., Blank, D., & Niemann, R. (2009). A hybrid index structure for geo-textual searches. *CIKM*, (S. 1625 - 1628).
Hariharan, R., Hore, B., Li, C., & Mehrotra, S. (2007). Processing Spatial-Keyword (SK) Queries in Geographic Information Retrieval (GIR) Systems. *Proceedings of the 19th International Conference on Scientific and Statistical Database Management* , 16 -.
Helgert, T. (2009). Automatisierte Erkennung geografischer Referenzen in unstrukturiertem Text. Hof: University of Applied Sciences Hof.
Jones, C. B., Abdelmoty, A. I., Finch, D., & Gaihua, F. (2004). The spirit spatial search engine: Architecture, ontologies and spatial indexing. *In Proc. 3rd Int. Conf. on Geographic Information Science*, (S. 125 - 139).
Names, U. B. (5. 4 2011). *U.S. Board on Geographic Names*. Abgerufen am 5. 4 2011 von http://geonames.usgs.gov/
Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web.* Stanford University.
Porter, M. F. (1980). An algorithm for suffix stripping. *Program* (14), 130 - 137.
School of Computer Science, C. U. (9. 3 2011). *Spatially-Aware Information Retrieval on the Internet*. Abgerufen am 9. 3 2011 von www.geo-spirit.org
Rill, S. (2005). SARA Semi Automatic Research Archiv - Entwicklung eines Konzepts zur otsbezogenen Archivierung von Nachrichten und Fakten - Demonstration dieses Konzepts in Form eines Prototyps. Hof: University of Applied Sciences Hof.