Search for journalists: New York Times challenge report

Corrado Boscarino
Centrum Wiskunde & Informatica
Science Park, 123
1098 XG Amsterdam, The Netherlands
corrado@cwi.nl

Arjen P. de Vries*
Centrum Wiskunde & Informatica
Science Park, 123
1098 XG Amsterdam, The Netherlands
arjen@acm.org

Wouter Alink
Centrum Wiskunde & Informatica
Science Park, 123
1098 XG Amsterdam, The Netherlands
alink@spinque.com

ABSTRACT

We investigate how a user-centred design to search can improve the support of user tasks specific to journalism. Illustrated by example information needs, sampled from our own exploration of the New York Times annotated corpus, we demonstrate how domain specific notions rooted in a field theory of journalism can be transformed into effective search strategies. We present a method for search-context aware classification of authorities, witnesses, reporters and columnists. A first search strategy supports the journalistic task of investigating the trustworthiness of a news source, whereas the second search strategy supports assessments of the objectivity of an author. In principle, these strategies can exploit the semantic annotations in the corpus; however, based on our preliminary work with the corpus, we conclude that straightforward full-text search is still a crucial component of any effective search strategy, as only recent articles are annotated, and annotations are far from complete.

Keywords

journalism, faceted search, interactive IR

1. INTRODUCTION

> A rhetoric is a social invention. It arises out of a time and place, a peculiar social context, establishing for a period the conditions that make a peculiar kind of communication possible and then is altered or replaced by another scheme. [1]

The particular context where a textual document has been written, the audience an author appeals to, and the goals she wants to achieve shape argumentation and the rules a written text has to comply with in order to be considered for publication. Conversely, understanding the features of a rhetoric shades some light on the context of a document and on its correct interpretation by a reader.

Journalism is one peculiar discursive practice, which fits what, in an attempt to establish a theory of journalism, Andrew R. Cline referred to as an *epistemological field* [2]. This model for the domain of journalism contains a characterisation of 1) what can and cannot be known, 2) the nature of the knower, 3) the nature of the relationships among the knower, the known, and the audience, and 4) the nature of language. Legitimate questions within this domain are "how to correctly represent a fact?", "why should I trust a certain source?" or "who's opinion does a certain text represents?". We believe that the richness of the semantics provided by the New York Times (NYT) annotated collection allows to specify search strategies in these domain specific terms, which are highly abstract from an information system perspective, but nevertheless most familiar to our target end user: a journalist in the process of writing an article. This is the main intuition behind our contribution to the New York Times challenge organised at HCIR2010.

The collection as a whole can be thought of as an implicit definition of a dominant journalistic field: through a careful process of editing and verification only an article that complies with all the requirements of this domain will appear in the newspaper. Our aim is to support a journalist in accessing the NYT collection by means of domain specific concepts, providing a highly inclusive system, which is intuitive, effective and entertaining to use.

In order to demonstrate our approach, we focus on two of what perhaps are the most important elements in any theory of journalism, writers and sources: "it is the curious relationship between the reporter as knower and the source as knower that creates much of what we understand as journalism. The reporter shifts between the roles of knower and conduit of the known."[2]. We built a search engine to retrieve sets of documents which support an end user in charting the entangled relationship between authors of an article and their sources, showing that abstract concepts can be translated to possibly very complex search strategies.

In the next section, after briefly discussing how authors and sources are understood in the domain theory of journalism that inspired our design, we introduce four important typifications into which, according to this theory, authors and sources can be classified. An author can aim to produce journalistic knowledge and be a Reporter or to express a private opinion and be a Columnist. A source on the other hand, that is a person who is mentioned in an article as in-

*Also affiliated to the Delft University of Technology

HCIR 2010 New Brunswick USA

formed about the facts, can derive her trustworthiness on a given issue in force of the circumstance that she was on the scene when a news event happened, being for that event a Witness, or because she is an expert on that matter, being then an Authority. Experts on journalism claim that it is of paramount importance, in order to evaluate any document on a given issue, to know who wrote the document and how information has been gathered on the field.

Section 3 explains how the four typifications can be translated into partitions of the search space: evidence to support belonging to one of the four categories here above can be gathered by means of different search strategies fired onto the NYT annotated corpus. Each strategy, which can be interactively tuned by a user, allows these abstract categories to be mapped onto faceted search processes. Finally, the last section summarises the main conclusions with respect to the specific challenge requirements.

2. AUTHORS AND SOURCES

According to the theories of rhetoric [1] and of journalism [2] that we considered, journalistic knowledge relies on inductive reasoning upon most often only indirect experience of events. The goal of a journalist is to collect and present different and possibly incompatible views on how events have been, leaving to a reader the burden of interpretation. While a Reporter must at least in principle abstain herself from commenting on the facts, a Columnist is sought after just because of her opinions: while both Reporters and Columnists write about facts, a competent reader is able to discern whether the focus is on a description of a fact or on a description of its possible meanings. Both rhetoric and visual cues contribute to allow a reader to establish memberships to one or another category. While a more rigid article structure is a common feature of a Reporter's work and the position of an article within the printed newspaper may also be used to determine an author's status, it is most often an author's reputation that affords a reader to either believe in the author's impartiality of to let her concentrate on the author's personal view on the facts.

Since a writer's experience on a fact is mostly mediated by an interpretation given by a source, evaluation of what has been written heavily depends on whether a source can be trusted on a particular issue. Readers mostly rely on their own background knowledge in order to evaluate a source's trustworthiness. Authorities derive their legitimacy to speak about a certain topic by virtue of being member of an official organisation, of academic or social achievements, or because of a past demonstration of their skills. When a reader does not have the necessary prior knowledge, she will typically rely on the information provided by the author, in order to determine a source's trustworthiness.

This is often the case for Witnesses, who's competence scope does not exceed a particular event: since a reader does not generally have much prior information about an event she is reading about, otherwise she would probably skip the article, trustworthiness of a Witness depends on the guarantees that a writer can provide in a reader's eyes. Reporters and Columnists will be given different weights when deciding on the sources they are quoting. The relationship between authors and sources, once an interpreting reader is included into the picture may become increasingly complex.

The system that we propose aims to extend the background knowledge that a reader commonly employs to assess both authors and sources, by letting the semantic annotations provided with the collection act as additional cues, allowing an end user to still be competent in evaluating this complicated relationship between authors and sources onto the much larger scale of the entire NYT collection.

3. SEARCH STRATEGIES

This is the core part of this report. Here we explain how the treatment of the two concepts of authors and sources in the domain theory of journalism can be translated to search strategies and how the documents within the annotated NYT corpus jointly with the search strategies support a user in making sense of those concepts. We used both the Apache Solr[1] and the Spinque[2] search servers to test informally the applicability of our proposed approach. Solr represents a classic text retrieval case, where the newspaper archive can be searched by ranking the full-text of the articles on their content. Spinque's *Strategy Builder* is a prototype environment where search processes are divided into two phases: search strategy definition, and the actual search. Search strategies are visually defined data-flows consisting of query terms, documents and named entities. While not the topic of this paper, the probabilistic database back-end on which search strategies are executed provides the flexibility needed to allow full exploration of the data space spanned by articles and their semantic annotations.

We think the level of control provided by the strategy builder provides to a user very powerful primitives for exploratory search. In our approach there is no set of documents that can be thought of as the denotation of the high level concepts: meaning arise from the act of exploring the collection and defining a search strategy as well as from reading the retrieved documents. Since with Spinque there is, even for a less experienced user, a clear division of meaning making labour between a visual development of a search strategy, faceted browsing of (intermediate) results and strategy refining, we expect further work on this subject to be carried forth in the form of 'search by strategy' processes.

3.1 Reporters and Columnists

In a first search task we suppose that, possibly as part of another search process, a user, in order to make sense of some document (Δ) that she retrieved, wants to collect evidence in favour or against its author being likely to deliver journalistic knowledge or rather personal opinions, although possibly very well motivated. The following semantic annotations are relevant to this task:[3] `taxonomicClassifiers`, `columnName`, `featurePage`, `authorBiography`, `body`, `byline`, and `people`.

The set of all `taxonomicClassifiers` forms a directed graph within the space of the whole collection: each document can be thought of as occupying a particular node of the graph and therefore a document's classification $\mathbf{C}$ can be defined as a set of nodes that contain a certain document. The search strategy to perform this task is an interactive and iterative process starting with a user, who must select for the document Δ a partition of the classifications $\mathbf{C}$, `columnName` and `featurePage` that she would consider definitely supporting the assertion that articles with those characteristic have been written by either a Reporter or a Columnist.

[1] `http://lucene.apache.org/solr/`

[2] `http://www.spinque.com/`

[3] Unless otherwise specified, a use of a fixed font refers to the scheme for Solr that has been provided with the NYT collection.

The system should be instructed on how to deal with borderline cases, whether to exclude them from search or to consider these features as supporting both cases.

The `authorBiography` and `byline` fields are used to query the `body` and, respectively, the `people` fields of documents in the collection. The results can again be partitioned by classifications **C**, `columnName` and `featurePage` and, if deemed necessary, the search process can continue, by applying the same strategy to any of the documents in the result set. Typically one step only is sufficient to complete the task, which can also be repeated by modifying the search strategy in any of its components.

Example.

Is the article "CELEBRATION; Chicago" an Opinionist's work?[4]
Its classification **C** is:

> `<Top, Features; Features, Travel; Travel, Columns; Columns, Celebration; Travel, Sophisticated Traveler Magazine; Features, Magazine; Top, Classifieds; Classifieds, Job Market; Job Market, Job Categories; Job Categories, Hospitality; Hospitality, Restaurant and Travel; Travel, Guides; Guides, Destinations; Destinations, North America; North America, United States; United States, Illinois; Illinois, Chicago; Travel, Guides; Guides, Destinations; Destinations, North America; North America, United States>`

which contain elements from both partitions: the node `<Columns, Celebration>` is typical of a Columnist's contribution, while `<Job Market, Job Categories>` more of a Reporter and `<Illinois, Chicago>` is neutral.

Using the annotations `byline: By Stephen McCauley` and `authorBiography: Stephen McCauley's most recent novel is "True Enough" (Simon & Schuster).` to query the collection, allow to retrieve *'True Enough': Just So-So Stories* [5], a review of McCauley's novel *True Enough*, which has been published in the Sunday Book Review.

By examining this evidence a user can conclude that the original article Δ should be regarded as a type Columnist's work. When necessary this process can continue, by using the `<str name="authorBiography">Louis Bayard's most recent novel is "Endangered Species."</str>` as the input of a second iteration step. It is important to notice that our proposed system does not provide a direct answer to the high level question of whether or not Stephen McCauley should be considered a Columnist in this case, but only the means for a user to make sense of this situation. The datum of a certain person being named in an article, which is also a book's review, licenses the statement that Stephen McCauley should be considered a Columnist only upon an autonomous interpretation by a user, who decides that the book is not written by a journalist, but by a novelist.

3.2 Authorities and Witnesses

Designing search strategies for Authorities and Witnesses is slightly more complicated as they require to detect events first: an Authority should be trusted because there is a set of past events in which the same person served as a source of reliable information, while a Witness should be trusted because many articles published around a certain event count her as a source. In addition to what we already used in the previous case, the following semantic annotations are also relevant to this task: `taxonomicClassifiers`, `publicationDate`, `locations`, `dateline`, and `text`.

The set of all `taxonomicClassifiers` forms again a directed graph, which can be used in the same way as in the previous case to partition the search space. Documents are moreover thought of as occupying the spatio-temporal space **S** defined by the `locations`, `dateline` and `publicationDate` fields.

[4] `www.nytimes.com/2002/03/03/magazine/celebration-chicago.html`

[5] `www.nytimes.com/2001/08/05/books/review/05BAYARDTw.html`

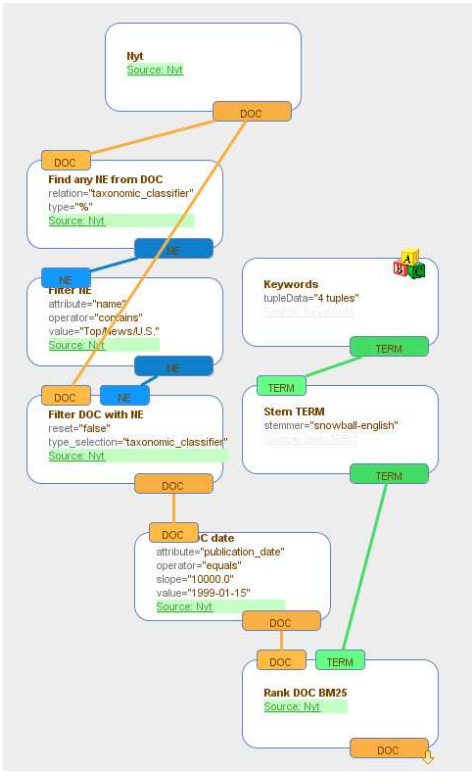

Figure 1: A Spinque strategy, consisting of connected building blocks, depicting a search approach taken to find the document set for Search A, in which the filter on date is a vague predicate. (no. results 30043)

A user may define an event for which she wants to determine which Authorities or Witnesses are potentially reliable source in the form of a set of documents, that, in order for the system to perform correctly, should occupy a narrow portion of the spatio-temporal space: most likely an event will be defined by onlya few documents, as we tested in these pilot experiments.

Provided that we can find an interval of area δ such that the intersection between a space **S** and a part of the taxonomy graph contains δ relevant documents about a given issue, while only $\frac{A}{\delta}$ relevant documents are outside the interval, for some positive constant A, we can define the same issue to denotate an event. Intuitively, many news articles have been published about the event around the same time and featuring the same places and names as the event: outside this local regularity the number of relevant documents decrease in measure of A.

Evidence for a person being a Witness can be presented to a user by collecting the documents which mention that person in their `text` fields and for which there are much less documents around other events. Conversely, evidence for Authorities, can be presented by collecting relevant document in more that one event.

Example.

Is justice Antonin Scalia an Authority? We first define an event as a non-empty partition of the search space (Search A):

> <keywords: newcomers state welfare policy>,<approx 1999-01-14>,<C=Top/News/U.S.>

contains amongst others 3 documents that rank high (using a custom Spinque search strategy) and are about the same event, see Fig.3.2:

- Supreme Court Hears Welfare Case (NYT, Jan 14, 1999)
- January 10-16; A New Look At the Right to Move (NYT, Jan 17, 1999)
- THE SUPREME COURT: CITIZENS' RIGHTS; Newcomers to States Have Right To Equal Welfare, Justices Rule (NYT, Jan 18, 1999)

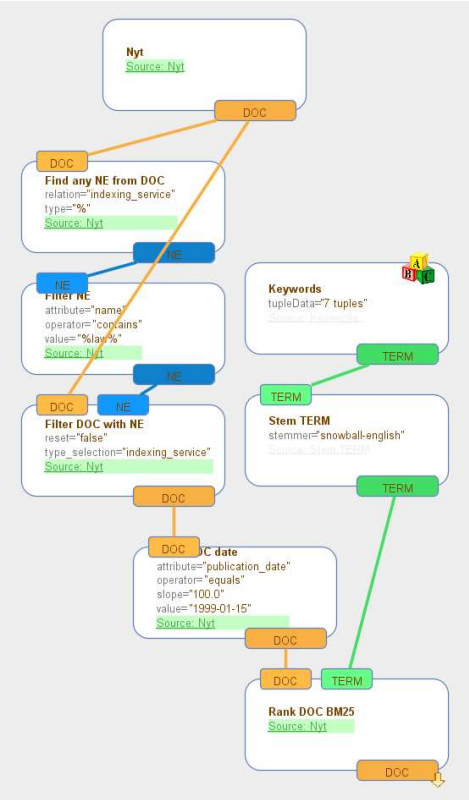

Figure 2: A Spinque strategy depicting a possible search approach taken to find the document set for Search B. (no. results: 16106.)

The same three documents could also have been found by a different search (using again a Spinque search strategy, Search B, see Fig.3.2):

> <keywords: individual rights equal citizen state supreme court>,<approx 1999-01-14>,<C≐Law>

Imagine the user of the system would have flagged these three documents, and would like to know more about their content. Two of these documents mention justice Antonin Scalia. The question of whether he is or not an authority on the issue depends on the search context. In the first search Antonin might not be regarded as an authority, as not many documents in the total result set (even if the date-filter is left out) are about him. He is not likely to be a Witness either as there are many documents about him outside this search result set. In the second search Antonin would very likely be an authority, as in many documents of the result set his name will be annotated, as he is a long serving supreme court member.

Notice that, again, we stress the importance of letting meaning arise from both an examination of the documents *and* from the search strategy that produced those documents: the event defined by the partition is a different event, albeit it contains the same documents as the previous one. When more events are generated in this way, adding multiple overlaps as in Fig.3.2, and upon examination of the evidence presented by the system, a user

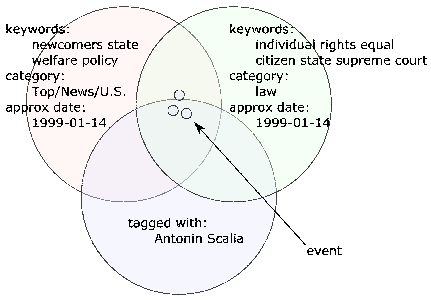

Figure 3: Events as overlapping partitions of the search space.

will probably conclude that the referent of `<str name="people"> Scalia, Antonin</str>` is an Authority, and possibly on the topic of 'Law and Legislation'. Notice that the amount of overlap of the second event can also trigger the conclusion that Antonin Scalia is also a Witness for that particular event. Albeit sentences of a court are generally easier events to detect, because they usually show a sharp temporal localisation, that is the day a sentence has been issued and a sharp spatial localisation as well, always Washington in case of the supreme court, we believe that this approach can be extended to more difficult cases where a user may require several steps, not only to establish membership to the classes of Authorities or Witness, but also to select appropriate events to initiate the task.

4. CONCLUSIONS

In this challenge report we explained how a system that allows end users to interactively map high level concepts to search strategies could be useful to make sense of those notions within a large and possibly unfamiliar repository. Motivated by an analysis of the available theories that have been developed by the same community to which our target users belong, we selected two abstract and deeply intertwined notions, that of author and source, that are difficult to approach using standard retrieval tools. The complexity of these notions and the open issue on whether or not any straightforward definition is possible or even desirable, calls for facilities to let users explore these notions, without taking an overly narrow stance on the issue.

We believe to have demonstrated the feasibility of our approach, meeting the main requirements of the challenge, for that we take advantage, when possible, of the extended semantic annotations, relying on text retrieval only when the annotations are unavailable of incomplete. The system we propose is effective mostly because the tasks are based on a domain model for exactly that particular class of users that we aim to support. It is also efficient, for that upon examination of only one set of documents a user is able to decide whether one of the two concepts apply. While guidance is still limited, as we do not yet provide any facility to determine how a modification in a search strategy affects its results, we claim to be successful in providing an application that is both transparent and fun to use. Because of the graphic interface of both the strategy builder and the graph exploration tool, which is currently under development, a user is able to identify which components and facets are being used at any moment and to very intuitively modify on the fly a search strategy.

5. REFERENCES

[1] J. A. Berlin. *Writing instruction in nineteenth-century American colleges / James A. Berlin ; with a foreword by Donald C. Stewart*. Southern Illinois University Press, Carbondale :, 1984.

[2] A. R. Cline. Toward a field theory of journalism, July 2010.