Documents Search Using Semantics Criteria

Santiago Cotelo Alejandro Makowski Luis Chiruzzo Dina Wonsever
Instituto de Computación
Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay
pln@fing.edu.uy

ABSTRACT
Current Information Retrieval systems generally search documents using a keywords model, which is often not expressive enough for the user. In this paper we describe some directions for improving an Information Retrieval system by letting the user specify different semantics constraints in her query, using a language based on a simplified version of first-order logic. The user can write queries that express the association between objects and attributes, temporal constraints and negation of attributes, and also perform synonyms expansion of queries. In order to evaluate the relevance of a candidate document with respect to the query, the dependency parse tree of the document is used, as well as other linguistic resources. The system was evaluated using a set of queries and a corpus extracted from the British newspaper *The Times*. The results are compared against the newspaper's own search engine and they look promising, showing an important improvement in precision in the first documents returned by the query.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *query formulation, retrieval models.*
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *language parsing and understanding*.

General Terms
Algorithms, Languages

Keywords
Information Retrieval; Natural Language Processing; dependency parsing; semantics; query language

1. INTRODUCTION
Current information retrieval (IR) systems frequently rely on a bag of words model. This means: the user specifies some words, and the engine looks for documents that contain one or more occurrences of those words. The set of documents that match the query are ordered according to different criteria (number of occurrences of each word, proximity of words, popularity of each document, etc.).

ESAIR'14, November 7, 2014, Shanghai, China.

ACM 978-1-4503-1365-0/14/11…$15.00
http://dx.doi.org/10.1145/2663712.2666187

The use of a simple bag of words model implies that the user will not be able to accurately specify which concepts she is looking for. For example, a user might enter the query *"american civilian and afghan solider"*, but as the word order does not matter in a bag of words, she might end up retrieving documents about *"american soldier and afghan civilian"*.

In order to overcome some of these limitations, we created a more expressive language for queries, which is based in a simplified version of first-order logic. Using this language, the user can specify different syntactic and semantic constraints so as to describe more precisely the documents she is looking for. The relevance of a document is scored using the dependency analysis of the document and other linguistic information.

In our analysis of the state of the art, we found that some research has been done in improving information retrieval systems by using dependencies information [1] [2] [3] [4] [5], though their work mainly refers to using dependency parsing to extract relevant information from natural language queries, and using this information to improve the scoring functions. We consider that the problem could be separated in two stages: first a user enters a query (keywords or free text) and the system transforms it into an intermediate representation; then the system uses this intermediate representation to classify the candidate documents as relevant or non-relevant. Although it is clear that both stages are necessary for a system that will be used by end users, in our work we are not considering the first stage, and we assume that our queries are issued using the intermediate representation, which we describe in the following section.

Classic information retrieval systems used a similar Boolean model: a query is a Boolean expression of terms. But the presence or absence of terms was considered in the whole document [6], so it was not possible to predicate over units smaller than a sentence. Another possibility is the use of proximity operators [7] but these cannot model long range dependencies.

2. THE QUERY LANGUAGE SEMQL
The language, which we call SemQL, provides a way to specify the following constraints:

- Association of objects and attributes: For example, being able to tell apart *"american civilian and afghan soldier"* from *"american soldier and afghan civilian"*. The two queries in our system would be different: the first would be *"american(x) ^ civilian(x) ^ afghan(y) ^ soldier(y)"* and the second would be *"american(x) ^ soldier(x) ^ afghan(y) ^ civilian(y)"*.
- Temporal semantics: Being able to specify a range of time in which an event happened. For example *"wars in the XIX century"*. In our system, this would be expressed as *"war(x)*

^ *between(x, 1801, 1900)"*, temporal expressions are in the TimeX3 [8] format.

- Negation of attributes: Being able to indicate that an attribute is not associated to a concept. For example, *"cats that are not white"*. In our system, this would be expressed as *"cat(x) ^ !white(x)"*.
- Synonyms expansion: Not a constraint per se, but the words in the queries are treated as concepts; and related concepts are also searched for. For example using the term *"battle"* also yields results where the term *"struggle"* is used. The related concepts are extracted from WordNet [9] [10].

Another constraint that was taken in consideration, but finally not implemented, is the ability to distinguish between the senses of a polysemic word. For example, when searching for *"apple"*, telling apart the fruit from the company.

The metaphor behind our query language is the following: a document mentions a set of concepts, which are represented by variables (*x, y*, etc.), and a query specifies a logic formula that must hold for those concepts in order to consider the document relevant. For instance, the query *"american(x) ^ civilian(x) ^ afghan(y) ^ soldier(y)"* is interpreted as: a relevant document contains a concept *x* which has the properties *"American"* and *"civilian"*, and a concept *y* which has the properties *"Afghan"* and *"soldier"*. In practice the scoring function also awards some points to documents that meet only part of the semantic constraints but not all of them.

The information retrieval system built contains a web search interface, a repository of indexed documents and a module that performs the relevance scoring for the documents. We used the open source Lucene + Solr [11] search platform to build our documents base and index. Each document is preprocessed by removing its stopwords and an inverted index is built using the document's keywords [12]. Furthermore, a full dependency parsing is performed for each document using the Stanford Parser [13] and the SUTime module [14] to extract the temporal expressions. A set of SemQL expressed concepts is extracted from this analysis and stored as metadata for the document, so as to speed up the scoring process afterwards.

When a user issues a query, the system first generates a set of keywords that are used to find a collection of candidate documents in Solr. Then the scoring module calculates the relevance of each document based on the query constraints and the SemQL expressions associated to the document. The result of this module is a score between 0 and 1 for each document. Each concept (variable) in the query is scored separately and then averaged. The set of rules we use for scoring a concept is shown in Table 1.

Table 1. Rules applied for scoring a document's concept against a query

Rule	Score
Object hit	+0.4
Attribute hit	+0.3 / #attributes
Temporal expression hit	+0.2 / #tempexpr
Explicit negation hit	+0.1 / #negations
Explicit negation miss	-0.35
Attribute miss	-0.25
Related term	*0.9
Difference in level	*0.2

3. RESULTS

In order to test the system, we created a corpus containing news articles from the UK newspaper The Times [15], using the newspaper's web search, which performs a standard keyword based search. The corpus was created by issuing a series of queries with variations on certain topics (*"high cost and low profits"*, *"afghan soldier american civilan"*, *"low blood pressure high cholesterol"*, and some combinations of those queries) and keeping the first 25 results for each query. The corpus has a total of 351 news articles.

The number of documents in the test corpus was not enough to perform a full formal evaluation of the system, but it helped us find some evidences about how the system works and which types of queries might yield better results.

With respect to association of objects and attributes, we compared how the systems perform when searching for documents about *"American civilans and Afghan soldiers"*. We use the query *"american(x) ^ civilian(x) ^ afghan(y) ^ soldier(y)"* for our system and the corresponding query *"american civilian afghan soldier"* in The Times search engine. Table 2 shows the first 25 documents retrieved by our system, and their corresponding ranking in The Times search engine. The first two results from our system have perfect score, and they contain the phrases *"as well as 14* ***Afghan soldiers*** *and one* ***American civilian****"* and *"two* ***American civilians*** *and two Afghans were killed in a shooting on an Afghanistan military base after an* ***Afghan soldier*** *opened fire yesterday"*. However, the first document is in position 18 of The Times results, and the second document is not present in the first 25 results. We consider that our system had better performance for this query, and it might indicate an improvement in precision over queries using only keywords.

We followed a similar process to test the behavior of synonyms expansion, temporal semantics and negation of attributes constraints. In the case of synonyms expansion, as expected, the recall of the queries was improved because relevant documents that contain terms related to the original query are retrieved. However as each term is considered separately, we lose the meaning of idioms and multiword expressions, which decreases precision. About temporal semantics, we found that it is possible to improve both precision and recall in some queries. The precision improves because documents with events that belong to the specified range are scored higher, and the recall improves because the search does not require an exact match of terms (e.g. the query specifies *"between(x, 2000, 2010)"* and the document contains the term *"2009"*). In the case of negation of attributes, we found that the precision of the queries was improved, because a lower score is assigned to the documents that contain attributes that are explicitly negated in the query.

As the size of the corpus was small and we could not perform many queries, these results cannot be considered a formal evaluation. They are meant to show the potential benefits of the system with respect to a basic keyword search.

Table 2. Results for query about American civilians and Afghan soldiers

Document Title	Our rank	*The Times* rank	Score
Sixteen Americans dead as helicopters crash in Afghan	1	18	1.0
Americans killed in shooting on Afghan base	2		1.0
Two US soldiers killed as Afghan teacher opens fire	3	12	0.85
Focus on how soldier was able to leave his base	4	6	0.85
Soldiers shot on Helmand guard duty were just days from	5	5	0.85
US military death toll in Afghanistan reaches 2,000	6	1	0.85
Survivors of Afghanistan massacre to give evidence	7		0.85
US troops shot dead after Ramadan meal invitation	8		0.85
Timeline: Attacks by rogue forces	9	24	0.78
Plan for Afghan exit as protests grow	10	22	0.78
Soldier could face death for Afghan 'killing spree'	11		0.78
Afghan retreat	12		0.78
Video shows US 'massacre' soldier hid weapon under shaw	13	7	0.70
US 'massacre' soldier flown out of Afghanistan	14	9	0.70
US forces fear reprisals over pictures of desecrated dead	15		0.70
Taleban vow revenge after US soldier shoots Afghan	16	13	0.70
US soldier on 3am rampage walked in and shot families	17	21	0.70
America insists it is on the path to success in Afghanistan	18		0.70
Staff Sergeant Robert Bales, accused of killing 16 Afghans	19	4	0.70
2,000th US soldier killed in Afghan green on blue attack	20	2	0.70
Real cost to the West could be a collapse of confidence	21	23	0.70
Crocker to step down as Ambassador to Afghanistan	22		0.70
US soldier Robert Bales accused of killing Afghans said	23	8	0.70
US Army drops charges against soldier in 'kill for sport'	24		0.63
You can't guard completely against these horrors	25		0.63

4. CONCLUSIONS AND FUTURE WORK

We built an information retrieval system that lets the user specify queries in a particular language based on a simplified version of first-order logic. Using this language the user might specify associations between objects and attributes in a document, temporal constraints associated to events and negation of attributes; also, the system performs synonyms expansions of the query terms. We use different linguistic resources in order to analyze which documents meet the specified constraints.

We conclude that, given that there are many mature linguistic resources that can be exploited (especially for English) it is possible and desirable to improve the current information retrieval methodology. Our results indicate that using the dependency parsing of documents (which could be preprocessed so as not to add excessive overhead at query time) it is possible to retrieve documents that are more relevant to the user.

In order to have a system that works as a whole, we need to build a module that transforms a standard keywords or free text query into our intermediate SemQL language. This module could also perform a dependency parsing, similar to the way we process the candidate documents.

5. REFERENCES

[1] T. Strzalkowski, G. Stein, G. Bowden Wise, P. Tapanainen, T. Jarvinen, A. Voutilainen, J. Karlgren, “Language Information Retrieval: TREC-7 Report”, 1998.

[2] C. Liu, H. Wang, S. McClean, E. Kapetianos, D. Carroll, “Weighting Common Syntactic Structures for Natural Language Based Information Retrieval”, Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 2010.

[3] J. Liu, P. Pasupat, Y. Wang, S. Cyphers, J. Glass, “Query Undersanting Enhanced by Hierarchical Parsing Structures”, 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2013.

[4] J. H. Park, W. B. Croft, “Query Term Ranking based on Dependency Parsing of Verbose Queries”, Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval Pages 829-830, 2010.

[5] K. T. Maxwell, J. Oberlander, W. B. Croft, “Feature-Based Selection of Dependency Paths in Ad Hoc Information Retrieval”, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013, pp 507–516.

[6] C. D. Manning, P. Raghavan, H. Schütze, “Introduction to Information Retrieval”, Cambridge University Press, 2008, s. 1.1, p. 4.

[7] “CQL, Contextual Query Language”, Library of Congress, http://www.loc.gov/standards/sru/cql/, last accessed 29 March 2014.

[8] TIMEX3 Specification, “TimeML 1.2.1 A Formal Specification Language for Events and Temporal Expressions”, http://www.timeml.org/site/publications/timeMLdocs/timeml_1.2.1.html#timex3, last accessed 23 April 2014.

[9] G. A. Miller, “WordNet: A Lexical Database for English”, Communications of the ACM Vol. 38, 1995, No. 11: 39-41.

[10] C. Fellbaum, “WordNet: An Electronic Lexical Database”, Cambridge, MA: MIT Press, 1998.

[11] “Apache Lucene – Apache Solr”, The Apache Software Foundation, http://lucene.apache.org/solr/, last accessed 20 April 2014.

[12] C. D. Manning, P. Raghavan, H. Schütze, “Introduction to Information Retrieval”, Cambridge University Press, 2008, s. 1.2, pp. 6-9.

[13] M. C. de Marneffe, B. MacCartney, C. D. Manning, 2006, “Generating Typed Dependency Parses from Phrase Structure Parses”, LREC 2006.

[14] A. X. Chang, C. D. Manning, 2012, “SUTIME: A Library for Recognizing and Normalizing Time Expressions”, 8th International Conference on Language Resources and Evaluation, LREC 2012.

[15] “The Times”, Times Newspaper Limited 2014, http://www.thetimes.co.uk, last accessed 20 April 2014.