Word Proximity Constraints: Information Retrieval meets Temporal Reasoning

Manolis Koubarakis
Intelligent Systems Laboratory
Dept. of Electronic and Computer Engineering
Technical University of Crete
73100 Chania, Crete, Greece
http://www.intelligence.tuc.gr/~manolis

Abstract

We study the data models $\mathcal{WP}$ and $\mathcal{AWP}$ that have been widely used for many years in the area of Information Retrieval. $\mathcal{WP}$ and $\mathcal{AWP}$ can be used to represent and query textual information under the Boolean model using the concepts of attributes with values of type text, and word proximity constraints. Variations of $\mathcal{WP}$ and $\mathcal{AWP}$ are in use in most deployed digital libraries using the Boolean model, text extenders for relational database systems (e.g., Oracle Text) and the search engine Altavista. We present the syntax, semantics and model theory of $\mathcal{WP}$ and $\mathcal{AWP}$ and analyze the complexity of query satisfiability and entailment. Since word proximity constraints are very similar to temporal constraints, the techniques we use in our analysis are similar to the ones developed in previous work on first-order theories of temporal constraints and temporal constraint databases.

1. Introduction

We revisit the data models $\mathcal{WP}$ and $\mathcal{AWP}$ that have been widely used for many years in the area of Information Retrieval (IR) [3]. The acronyms $\mathcal{WP}$ and $\mathcal{AWP}$ were introduced by us in [13, 10].

Data model $\mathcal{WP}$ is based on *free text* and its query language is based on the *boolean model with word proximity constraints*. Data model $\mathcal{AWP}$ extends $\mathcal{WP}$ and it is based on *attributes* with free text as values. The query language of $\mathcal{AWP}$ is a simple extension of the query language of $\mathcal{WP}$ so that attributes are included.

In the models $\mathcal{WP}$ and $\mathcal{AWP}$ *word patterns* of the form $w_1 \prec_{[l,u]} w_2$ stand for "word w_1 is *before* w_2 and is separated by w_2 by *at least* l and *at most* u *words*". For example, $luxurious \prec_{[0,3]} hotel$ denotes that the word "hotel" appears before word "luxurious" and at a distance of at least 0 and at most 3 words. The word pattern $Holiday \prec_{[0,0]} Inn$ denotes that the word "Holiday" appears exactly before word "Inn" so this is a way to encode the phrase "Holiday Inn".

Word patterns were originally introduced in the area of Information Retrieval and have been implemented in many digital library systems in wide use today. Word patterns in IR systems encode word proximity *constraints* (in the sense of constraint databases [9]) using *proximity operators* kW and kN where k is a natural number [3]. The word pattern $w_1\ kW\ w_2$ stands for "word w_1 is before w_2 and is separated by w_2 by at most k words". In our work this can be captured by $w_1 \prec_{[0,k]} w_2$. The operator kN is used to denote distance of at most k words where the order of the involved patterns does not matter. In our framework the expression $w_1\ kN\ w_2$ can be written as $w_1 \prec_{[0,k]} w_2 \vee w_2 \prec_{[0,k]} w_1$. Strangely enough proximity operators are not popular with current search engines although they could offer a useful extension of "phrase search" facilities. From the well-known search engines only Altavista (`www.altavista.com`) has an operator $NEAR$ which means word-distance 10. There are also advanced IR models such as the model of proximal nodes [15] with proximity operators between arbitrary structural components of a document (e.g., paragraphs or sections).

In the database literature proximity operators have been studied by Chang and colleagues in the context of integrating heterogeneous digital libraries [4, 5, 6]. To the best of our knowledge, these papers contain the only comprehensive treatment of proximity operators that exists in the literature (including IR papers). Our works owes a lot to [4, 5, 6]: the model $\mathcal{AWP}$ is essentially the model of [4, 5, 6] but with a *different* class of word patterns.

We have recently used the model $\mathcal{AWP}$ for representing and querying resource meta-data in the distributed information alert system DIAS [10] and the peer-to-peer system P2P-DIET (`http://www.intelligence.tuc.gr/p2pdiet`) [12, 8, 7].

Current extensions of relational database products such as Oracle Text aimed at the support of information retrieval and filtering applications offer full support for proximity operators [1]. Database systems for XML will probably be the next place where we will encounter them. We expect the recent W3C working draft [14] and papers like [2] to pave the way for the introduction of such IR features in XML query languages XQuery and XPath.

In our talk we will present the syntax, semantics and model theory of $\mathcal{WP}$ and $\mathcal{AWP}$ and analyze the complexity of query satisfiability and entailment. Since word proximity constraints are very similar to *temporal constraints*, the techniques we use in our analysis are similar to the ones developed in previous work on first-order theories of temporal constraints and temporal constraint databases. We hope that our work (as presented in detail in [10, 10, 16, 11] and summarized in our presentation at the workshop) will be interesting to temporal reasoning researchers since we demonstrate new problems where techniques from temporal reasoning can be usefully applied.

References

[1] Oracle Text Reference Guide. Available at Oracle Text Home Page http://otn.oracle.com/products/text.

[2] S. Amer-Yahia, M. F. Fernandez, D. Srivastava, and Y. Xu. Phrase Matching in XML. In *Proceedings of the 29th International Conference on Very Large Databases (VLDB)*, pages 177–188, Berlin, Germany, September 2003.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[4] C.-C. K. Chang, H. Garcia-Molina, and A. Paepcke. Boolean Query Mapping across Heterogeneous Information Sources. *IEEE Transactions on Knowledge and Data Engineering*, 8(4):515–521, 1996.

[5] C.-C. K. Chang, H. Garcia-Molina, and A. Paepcke. Predicate Rewriting for Translating Boolean Queries in a Heterogeneous Information System. *ACM Transactions on Information Systems*, 17(1):1–39, 1999.

[6] K. C.-C. Chang. *Query and Data Mapping Across Heterogeneous Information Sources*. PhD thesis, Stanford University, January 2001.

[7] S. Idreos, M. Koubarakis, and C. Tryfonopoulos. P2P-DIET: Ad-hoc and Continuous Queries in Super-peer Networks. In *Proceedings of the IX International Conference on Extending Database Technology (EDBT04)*, pages 851–853, Heraklion, Crete, Greece, 14–18 March 2004.

[8] S. Idreos, C. Tryfonopoulos, M. Koubarakis, and Y. Drougas. Query Processing in Super-Peer Networks with Languages Based on Information Retrieval: the P2P-DIET Approach. In *Proceedings of the 1st International Workshop on Peer-to-peer Computing and Databases (in conjunction with EDBT'04)*, March 2004.

[9] P. Kanellakis, G. Kuper, and P. Revesz. Constraint Query Languages. In *Proceedings of the 9th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 299–313, 1990.

[10] M. Koubarakis, T. Koutris, C. Tryfonopoulos, and P. Raftopoulou. Information Alert in Distributed Digital Libraries: The Models, Languages and Architecture of DIAS. In *Proceedings of the 6th European Conference on Digital Libraries (ECDL2002)*, volume 2458 of *LNCS*, pages 527–542, September 2002.

[11] M. Koubarakis, S. Skiadopoulos, and C. Tryfonopoulos. Complexity Results for Boolean Information Retrieval Languages Based on Attributes and Word Proximity Constraints. Unpublished Manuscript.

[12] M. Koubarakis, C. Tryfonopoulos, S. Idreos, and Y. Drougas. Selective Information Dissemination in P2P Networks: Problems and Solutions. *ACM SIGMOD Record, Special issue on Peer-to-Peer Data Management, K. Aberer (editor)*, 32(3), September 2003.

[13] M. Koubarakis, C. Tryfonopoulos, P. Raftopoulou, and T. Koutris. Data models and languages for agent-based textual information dissemination. In *Proceedings of the 6th International Workshop on Cooperative Information Agents (CIA2002)*, volume 2446 of *LNAI*, pages 179–193, September 2002.

[14] XQuery and XPath Full-Text Use Cases. W3C Working Draft 14 February 2003. Available at `http://www.w3.org/TR/xmlquery-full-text-use-cases`.

[15] G. Navarro and R. A. Baeza-Yates. Proximal Nodes: A Model to Query Document Databases by Content and Structure. *ACM Transactions on Information Systems*, 15(4):400–435, 1997.

[16] C. Tryfonopoulos, M. Koubarakis, and Y. Drougas. Filtering Algorithms for Information Retrieval Models with Named Attributes and Proximity Operators. In *Proceedings of the 27th Annual ACM SIGIR Conference*, Sheffield, United Kingdom, July 25-July 29 2004. Forthcoming.