Exploiting Inference from Semantic Annotations for Information Retrieval

Reflections from Medical IR

Guido Zuccon[1], Bevan Koopman[2,1], and Peter Bruza[1]
[1]School of Information Systems, Queensland University of Technology, Australia
[2]Australian E-Health Research Centre, CSIRO, Australia
{g.zuccon, b.bruza}@qut.edu.au, bevan.koopman@csiro.au

ABSTRACT

The increasing amount of information that is annotated against standardised semantic resources offers opportunities to incorporate sophisticated levels of reasoning, or inference, into the retrieval process. In this position paper, we reflect on the need to incorporate semantic inference into retrieval (in particular for medical information retrieval) as well as previous attempts that have been made so far with mixed success. Medical information retrieval is a fertile ground for testing inference mechanisms to augment retrieval. The medical domain offers a plethora of carefully curated, structured, semantic resources, along with well established entity extraction and linking tools, and search topics that intuitively require a number of different inferential processes (e.g., conceptual similarity, conceptual implication, etc.). We argue that integrating semantic inference in information retrieval has the potential to uncover a large amount of information that otherwise would be inaccessible; but inference is also risky and, if not used cautiously, can harm retrieval.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]
General Terms: Theory
Keywords: Semantic Annotations, Medical Information Retrieval, Inference

1. INTRODUCTION

Advances in 'entity linking', as well as the increasing practice of augmenting web content with embedded structured data, has resulted in an increasing interest in leveraging this semantic information for more effective information retrieval (IR). While the problem of entity retrieval has been well studied (e.g., [3, 9]), the exploitation of entities or concepts[1] linking and semantic annotations is still an open area of research, as witnessed by the Exploiting Semantic Annotations in Information Retrieval (ESAIR) workshops [1, 11].

The availability of semantic resources like knowledge bases, where information is structured and linked, as well as mapping tools, that allow for individuating entities and link them to the knowledge resource, offers the possibility of understanding the meaning and reason about information. We speculate that the incorporation of the reasoning process, or inference, within the retrieval process can generate a breakthrough in search engine technologies, providing methods that go beyond the simple matching of keywords (or even entities) offered by today's technologies; similar speculations have been made by others in the past, e.g., [17]. These methods may not only lead to improved retrieve effectiveness, but to more precise and articulated query statements, more informative result analytics and more grounded handling for faceted and exploratory search.

In this paper, we briefly discuss attempts to integrate inference mechanisms in information retrieval, both when using semantic annotations that map to structured or semi-structured knowledge resources and when using a weaker form of semantic relation (e.g., early work in information retrieval that instructed semantic relationships based solely on similarities between terms). We then examine recent findings and advances in medical IR that have attempted to integrate well structured semantic information within the retrieval process: this domain offers well established, manually curated knowledge resources and complex information needs that require inferential retrieval solutions. Observations from this domain reveal that integrating inference within the retrieval process shows promise but also exhibits some shortcomings, thus constituting an interesting open area of research.

2. INFERENCE IN INFORMATION RETRIEVAL

The integration of inference (and semantic inference in particular) has been the focus of a large body of IR research. There have been various usages and notions of inference in IR (and in general), ranging from logical, theoretical models, to textual entailment, to vague associative matching.

The introduction of logical models for IR, and in particular of the Logical Uncertainty Principle (LUP) [18], proposed to evaluate the relevance of documents as the extent to which propositions representing queries and those representing documents could be logically implied (i.e., $P(q \rightarrow d)$ or $P(d \rightarrow q)$, depending on the model). One such example is the Logical Imaging technique (LI) [7, 23], where the truth of the logical implication is evaluated as a function of the

[1]In the following we shall use concept and entity interchangeably.

ESAIR'14, November 07 2014, Shanghai, China

http://dx.doi.org/10.1145/2663712.2666197.

expected mutual information between terms. Although rudimentary in the way they capture semantics (and semantic relations), in particular for their focus on terms rather than entities, these techniques have demonstrated no significant performance gains [24] over more traditional (not semantical and not inferential) keyword-based matching strategies. Similar theoretical formulations aimed at further exposing inference mechanisms within the retrieval process have also not accomplished much success, e.g., [20]. Alternative inference mechanisms (although not relying on structured semantic information), like information flow [5], have been used to produce better query models for improving query and user understanding, as well as retrieval, but have not been subject of much uptake in the IR community.

Graph-based models, like Turtle and Croft's [17] inference network, have also been used to define inference mechanisms to augment the retrieval process and form the basis of the successful Indri query language [15], which allows complex structured queries to be constructed and evaluated, although not relying on structured semantic information. Extensions of this system have progressively increased the amount of semantic inference achievable (e.g., [4]), and the recent work of Dalton et al. [8] demonstrates how query representations can be enriched with features from semantic annotations and their links to knowledge bases[2].

3. INFERENCE IN MEDICAL INFORMATION RETRIEVAL

The medical domain has well crafted, robust, extensive structured semantic resources like the UMLS and SNOMED CT, among others, as well as established tools that map free-text to such structured semantic resources, e.g., Metamap [2]. Velupillai already addressed the ESAIR audience about the potential benefit that semantic annotations have in the clinical domain, in particular for hypothesis generation, adverse event identification and patient similarity [19]. Recent studies have furthermore suggested that medical information retrieval is characterised by problems that are *inherently inferential* and thus, integrating inference within the retrieval process may provide significant breakthroughs in the way information is searched, analysed and reported in this domain. For example, Koopman and Zuccon [13] analysed some of the reasoning that clinical users employ when assessing the relevance of clinical documents to queries aimed at identifying patient cohorts that satisfy a number of inclusion criteria for clinical trial. They showed that users apply inference mechanisms to determine temporality, the relations between query aspects, cause-effect relations, etc. Similarly, Koopman [12] performed a meta-analysis of the different factors that require semantic inference to augment retrieval in the medical domain; they identified the following as the primary factors: (1) semantic similarity and (2) granularity, (3) conceptual implication, (4) temporality, (5) level of uncertainty, (6) negation and contextually, (7) dependencies between entities.

Rudimentary inference mechanisms were used in a heuristic-driven model that exploited semantic concepts and is-a (a.k.a. subsumption) relations [25], achieving a form of semantic

[2]Note that in this brief overview of inference for information retrieval we have only considered some of the formal models developed in the IR community and have not provided an account for alternative, less formal approaches arising from the Semantic Web community, e.g., [10].

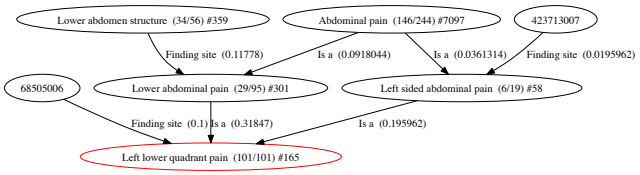

Figure 1: Partial traversal graph obtained by the Koopman's GIN model for query 147 of the TREC MedTrack collection: red nodes represent query concepts, black nodes concepts that are involved in the retrieval process thanks to the inference mechanism. Image courtesy of Koopman [12].

query expansion that was driven by the inference that, for example, a sought concept was the parent of a more specific concept present in a potentially relevant document. Cohen et al. [6] has recently demonstrated the use of more sophisticated levels of semantic inference accomplished through analogical reasoning based on high-dimensional vector representations to drive pseudo-relevance feedback. Although this method does encode higher level semantic inferences it does not provide substantial gains in terms of retrieval effectiveness and is indeed inferior to well engineered statistical approaches to search. A full-fledged semantic inference model for medical information retrieval has been proposed by Koopman [12] (the Graph Inference Model, GIN). The model, partially inspired by some of the literature reviewed in section 2 (LUP, LI and Inference networks), performs an implicit query expansion at retrieval time by traversing a graph-based semantic representation of the corpora, where both queries and documents are annotated with semantic concepts (forming the nodes in the graph) and edges encode different semantic relations between concepts (e.g., is-a, finding site, active ingredient, etc.). An example of the inference mechanism used by the GIN is given in Figure 1. The figure highlights how concepts that are inferred from the query concepts (by exploiting semantic relations directly encoded in the underling knowledge base) permit to retrieve more relevant documents than what the original query concepts retrieved in first instance. (In Figure 1: 56, 244, 95 and 19 more relevant documents for each of the concepts considered by the inference mechanism.) Thorough empirical experimentation however demonstrated that semantic inference, as encoded in the GIN, is a risky mechanism that, if not used cautiously, can harm retrieval [12]. Similar conclusions were observed in the work of Cohen et al. [6]. An important observation springing from Koopman's work is that it is the quality of the structured knowledge resource that influences the quality of inferences used for retrieval: in particular, inferences that may be logically valid from a representational perspective may not provide valuable information when used for retrieval.

Despite the previously mentioned efforts (among others) to integrate inference mechanisms within retrieval models for medical search, the most successful approaches in this domain used semantic annotations, but only with weak inference mechanisms, by combining structured domain knowledge with document and corpus statistics. For example, using concepts, semantic types (higher level groupings of concepts, e.g., diseases, organisms) and corpus statistics, Zhou et al.[21] were able to derive implicit relations between con-

cepts, which could be used for query expansion; this was the best approach at the TREC Genomics Track [22]. Similarly, recent work by Limsopatham et al. [14] partially exploits semantic dependency information between concepts, but relying on a powerful statistical diversification approach for document ranking adapted from web retrieval [16] rather than more semantically-grounded inference mechanisms.

4. CONCLUSIONS

This paper presented a brief and non-exhaustive account of attempts to integrate (semantic) inference mechanisms in information retrieval, and in particular in the medical domain. Medical IR offers a fertile playground where to develop and evaluate semantic-based retrieval techniques that go beyond the matching of semantic entities mentioned in queries, performing semantic inference using the structured (or semi-structured) information encoded in domain-specific knowledge bases. Based on findings from medical IR, we have argued for the need for integrating semantic inference into the retrieval process and we have shown that this has the potential, for example, to uncover a large amount of information that otherwise would be inaccessible. However, empirical results obtained in the medical domain have also demonstrated that fully embracing inference mechanisms can harm retrieval and that weaker and more conservative semantic inference mechanisms, combined with popular statistical methods are capable of robustly improve retrieval effectiveness, although not providing a full account of the required inference types.

Acknowledgements. The authors would like to acknowledge Reviewer 1 for the insightful commentary and pointers.

5. REFERENCES

[1] O. Alonso and H. Zaragoza. Exploiting semantic annotations in information retrieval. In *Advances in Information Retrieval*, pages 712–712. Springer, 2008.

[2] A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proc. of AMIA*, page 17. American Medical Informatics Association, 2001.

[3] K. Balog, P. Serdyukov, and A. P. d. Vries. Overview of the trec 2010 entity track. In *Proc. of TREC 2010*, 2010.

[4] M. Bendersky, D. Metzler, and W. B. Croft. Learning concept importance using a weighted dependence model. In *Proc. of WSDM'10*, pages 31–40. ACM, 2010.

[5] P. D. Bruza and D. Song. Inferring query models by computing information flow. In *Proc. of CIKM'02*, pages 260–269. ACM, 2002.

[6] T. Cohen, D. Widdows, and T. Rindflesch. Expansion-by-analogy: A vector symbolic approach to semantic search. In *Proc. of QI'14*, 2014 (to appear).

[7] F. Crestani and C. J. van Rijsbergen. Information retrieval by logical imaging. *Journal of Documentation*, 51(1):3–17, 1995.

[8] J. Dalton, L. Dietz, and J. Allan. Entity query feature expansion using knowledge base links. In *Proc. of SIGIR'14*, SIGIR '14, pages 365–374, New York, NY, USA, 2014. ACM.

[9] G. Demartini, T. Iofciu, and A. P. De Vries. Overview of the inex 2009 entity ranking track. In *Focused Retrieval and Evaluation*, pages 254–264. Springer, 2010.

[10] T. Finin, J. Mayfield, A. Joshi, R. S. Cost, and C. Fink. Information retrieval and the semantic web. In *Proc. of HICSS'05*, 2005.

[11] J. Kamps, J. Karlgren, P. Mika, and V. Murdock. Fifth workshop on exploiting semantic annotations in information retrieval: Esair"12). In *Proc. of CIKM'12*, pages 2772–2773. ACM, 2012.

[12] B. Koopman. *Semantic Search as Inference: Applications in Health Informatics*. PhD thesis, School of Information Systems, Queensland University of Technology, 2014.

[13] B. Koopman and G. Zuccon. Why assessing relevance in medical ir is demanding. In *MedIR 2014*, 2014.

[14] N. Limsopatham, C. Macdonald, and I. Ounis. Modelling relevance towards multiple inclusion criteria when ranking patients. In *Proc. of CIKM'14*, 2014 (to appear).

[15] D. Metzler and W. B. Croft. Combining the language model and inference network approaches to retrieval. *IP&M*, 40(5):735–750, 2004.

[16] R. L. Santos, C. Macdonald, and I. Ounis. Exploiting query reformulations for web search result diversification. In *Proc. of WWW'10*, pages 881–890. ACM, 2010.

[17] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM TOIS*, 9(3):187–222, 1991.

[18] C. J. Van Rijsbergen. A non-classical logic for information retrieval. *The computer journal*, 29(6):481–485, 1986.

[19] S. Velupillai. Semantic annotations in clinical documentation: Exploring potentials for future information retrieval. In *Proc. ESAIR'10*, pages 9–10. ACM, 2010.

[20] S. K. M. Wong and Y. Y. Yao. On modeling information retrieval with probabilistic inference. *ACM TOIS*, 13(1):38–68, 1995.

[21] W. Zhou, C. Yu, N. Smalheiser, V. Torvik, and J. Hong. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *Proc. of SIGIR'07*, pages 655–662, Amsterdam, The Netherlands, 2007.

[22] W. Zhou, C. Yu, V. Torvik, and N. Smalheiser. A concept-based framework for passage retrieval in genomics. In *Proc. of TREC'06*, pages 14–17, 2006.

[23] G. Zuccon, L. Azzopardi, and C. van Rijsbergen. A formalization of logical imaging for information retrieval using quantum theory. In *Proc. of DEXA'08*, pages 3–8. IEEE, 2008.

[24] G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen. Revisiting logical imaging for information retrieval. In *Proc. of SIGIR'09*, pages 766–767. ACM, 2009.

[25] G. Zuccon, B. Koopman, A. Nguyen, D. Vickers, and L. Butt. Exploiting medical hierarchies for concept-based information retrieval. In *Proc. of ADCS'12*, pages 111–114. ACM, 2012.