Noname manuscript No.
(will be inserted by the editor)

A Survey of Approaches for Ranking on Structured Data

First Author · Second Author

Received: date / Accepted: date

Abstract Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed.

Keywords First keyword · Second keyword · More

1 Style

Text with citations [**?**] and [**?**].

1.1 Subsection title

as required. Don't forget to give each section and subsection a unique label (see Sect. 1).

Paragraph headings Use paragraph headings as needed.

$$a^2 + b^2 = c^2 \tag{1}$$

F. Author
first address
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: fauthor@example.com

S. Author
second address

Fig. 1: Please write your figure caption here

Table 1: Please write your table caption here

first	second	third
number	number	number
number	number	number

2 Introduction

Outline

3 The Ranking Problem

3.1 Ranking on Structured Data

Structured data: can be conceived as graphs
Ranking: traditionally, core IR problem, many techniques have been adopted to deal with structured data
Ranking over structured data: ranked list of results
Types of ranked results: entities, relations, subgraphs, entire datasets (graphs)
Evaluating ranked results: different metrics

3.2 Ranking on Structured Data vs. Ranking on Unstructured Data

Discuss differences
Say we focus on structured data

3.3 Ranking vs. Matching

Discuss differences
Say we focus on ranking

3.4 Applications

Optional / can be short but should provide good coverage of different application domains

Fig. 2: Please write your figure caption here

4 Ranking Approaches

4.1 Historical Development

4.2 Generic Architecture

4.3 Main Dimensions

Data: structured data conceived as graphs? All existing approaches are applicable to data graphs or are there any that exploit special characteristics of specific types of structured data? Only in the latter case, approaches shall be distinguished along this data dimension

Queries: same as data dimension: are there many different approaches dealing with different types of queries so that we can use query as a dimension to distinguish approaches?

Features: content, structure, context, two meanings of context: context of user; and context of the data such as trust values of source, truth value / confidence degree of individual records; make clear this survey does not discuss user-context based methods in details

Results: entity, relationships, subgraphs, graphs (entire dataset)

Techniques:
- may vary in terms of methods: used *NLP* for understanding keywords, used *ML* for classifying keywords (I know some ML-based query classification approaches for document retrieval, are there also examples for structured data ranking?) and for learning to rank, (other) statistics-based methods: Vector Space Model (VSM), Language Modeling (LM), Information Theory
- may vary in terms of heuristics: two main ones are query-relevance and popularity; other heuristics are proximity, informativeness (based on Information Theory, e.g. entropy) and context-based: trust, truth value, locality etc.

4.4 Foundation

Here we discuss all basics needed to understand the approaches presented in the following sections.
Vector-Space Model: also discuss pivoted normalization and point out problems of short document length etc. in the context of structured data
Language Modeling: also discuss smoothing strategies
Link Analysis
Learning to Rank
...

4.5 Taxonomy of Ranking Approaches

Classify approaches mainly based on the type of *heuristics* they used, i.e. (1) Query-relevance, (2) Popularity, (3) Other Heuristics and (4) All Heuristics / LTR.

Those that use same heuristics are distinguished in terms of *methods*. E.g. query-relevance based solutions can be further distinguished in terms of VSM-based and LM-based approaches.

an

5 Query-relevance

6 Popularity

Wikipedia defines popularity[1] as the quality of being well-liked or common, or having a high social status. Translating this to information systems, popularity ranking algorithms rely on link analysis to compute the relevance of a node of the graph in terms of its input and output linkage. With the arrival of the Web, link analysis was proposed as a new method to rank the hypertext documents considering the way the information is represented and related [](Getoor and Diehl, 2005). Unlike content-based (statistics-based?) ranking, link analysis strategies try to incorporate structural features of the information during the ranking computation. Link analysis is a technique relying on examining the graph structure established among items, where the nodes of the graph are the items to rank and the edges are the relationships or links among items. By inspecting the graph structure implicit properties can be derived and included in the ranking process. Link analysis can be thought as a case of success, which was originally implemented in algorithms like PageRank [](Brin & Page, 1998), HITS [](Kleinberg, 1998) and SALSA [](Lempel & Moran, 2000), commercially exploited by the most popular search engines. Inspired by this philosophy several extensions have been developed that increase the corpus of link analysis methodologies to deal with structured and semi-structured information[2], namely:

[1] http://en.wikipedia.org/wiki/Popularity

[2] This classification has been taken from [](Delbru, Toupikov, Catasta, Tummarello & Decker, 2010)

- Weighted link analysis. The aim of this technique is to assign more relevance to certain kind of links depending on its type during the ranking computation. A major challenge is how to assign the weight to the links without having negative performance implications. Most of the approaches under this classification were proposed in the topic of database research, and for this reason they are not directly applicable on web-scale. As an example for this category we can consider the works described in [](Xing & Ghorbani, 2004) and [](Baeza-Yates & Davis, 2004).
- Hierarchical link analysis. This technique performs a layered exploration of the underlying data and it is intended for distributed environments. For example, first considering relationships among super nodes or datasets and secondly considering relationships among resources. An example for this category can be found in [](Xue, et al. 2005).
- Semantic Web link analysis. This family of methods tries to exploit the semantic of relationships during the ranking process. This technique can be thought as an evolution of the weighted link analysis applied to the Semantic Web context. As an example we can consider the algorithms described in [](Ding et al., 2004) and [](Anyanwu, Maduko, & Sheth, 2005), being the last one the culmination of the ideas previously introduced in [](Anyanwu & Sheth, 2002) and [](Anyanwu & Sheth, 2003). Examples of early commercial applications exploiting semantic Web link analysis include the works described in [](Sheth, Avant & Bertram, 2001), [](Avant et. al, 2002) and [](Sheth, 2005).

7 Other Heuristics

8 Combining heuristics

Hybrid heuristics vs composite heuristics

Within composite heuristics can find manual composition vs automatic composition (i.e. Learning to rank). This taxonomy is similar to the example pointed by Tran about matching approaches.

9 Sample Approaches from the Literature

Before, we discuss the concepts/techniques behind existing approaches. Here we provide a sample

10 Open Research Challenges

11 Conclusion

For consistency, we use bibtex entries from dblp, e.g. `http://www.dblp.org/rec/bibtex/conf/sigmod/BalminKT12`. Many bibtex entries are already in the paper.tex file so just searched for these entries there first and add only new entries when needed to avoid redundancy.

References