Web Semantics: Science, Services and Agents on the World Wide Web 30 (2015) 82–105

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

journal homepage: www.elsevier.com/locate/websem

An overview of semantic search evaluation initiatives

Khadija M. Elbedweihy [a,*], Stuart N. Wrigley [a], Paul Clough [b], Fabio Ciravegna [a]

[a] Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, UK

[b] Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, UK

ARTICLE INFO

Article history:
Received 1 July 2013
Received in revised form
10 October 2014
Accepted 13 October 2014
Available online 22 October 2014

Keywords:
Semantic search
Usability
Evaluation
Benchmarking
Performance
Information retrieval

ABSTRACT

Recent work on searching the Semantic Web has yielded a wide range of approaches with respect to the underlying search mechanisms, results management and presentation, and style of input. Each approach impacts upon the quality of the information retrieved and the user's experience of the search process. However, despite the wealth of experience accumulated from evaluating Information Retrieval (IR) systems, the evaluation of Semantic Web search systems has largely been developed in isolation from mainstream IR evaluation with a far less unified approach to the design of evaluation activities. This has led to slow progress and low interest when compared to other established evaluation series, such as TREC for IR or OAEI for Ontology Matching. In this paper, we review existing approaches to IR evaluation and analyse evaluation activities for Semantic Web search systems. Through a discussion of these, we identify their weaknesses and highlight the future need for a more comprehensive evaluation framework that addresses current limitations.

1. Introduction

The movement from the 'web of documents' towards structured and linked data has made significant progress in recent years. This can be witnessed by the continued increase in the amount of structured data available on the Web, as well as the work carried out by the W3C Semantic Web Education and Outreach (SWEO) Interest Group's community project *Linking Open Data*[1] to link various open datasets. This has provided tremendous opportunities for changing the way search is performed and there have been numerous efforts to exploit these opportunities in finding answers to a vast range of users' queries. These efforts include *Semantic Web search engines*, such as Swoogle [1] and Sindice [2], which act as gateways to locate Semantic Web documents and ontologies in a similar fashion to how Google and Yahoo! are used for conventional Web search. Whilst these systems are intended for Semantic Web experts and applications, another breed of tools has been developed to provide more accessible approaches to querying structured data. This includes *natural language interfaces* operating in a single domain, such as NLP-Reduce [3] and Querix [4], or in multiple and heterogeneous domains, such as PowerAqua [5] and Freya [6]; *view-based interfaces* allowing users to explore the search space whilst formulating their queries, such as K-Search [7] and Smeagol [8]; and *mashups*, integrating data from different sources to provide rich descriptions about Semantic Web objects, such as Sig.ma [9] and VisiNav [10].

Similar to designing and developing Information Retrieval (or search) systems more generally, evaluation is highly important as it enables the success of a search system to be quantified and measured [11]. This can involve evaluating characteristics of the IR system itself, such as its retrieval effectiveness, or assessing consumers' acceptance or satisfaction with the system [12]. For decades, the primary approach to IR evaluation has been system-oriented (or batch-mode), focusing on assessing how well a system can find documents of interest given a specification of the user's information need. One of the most used methodologies for conducting IR experimentation that can be repeated and conducted in a controlled lab-based setting is test collection-based evaluation [13,14,11,15]. Commonly known as the Cranfield methodology, this approach has its origin in experiments conducted at Cranfield library in the UK [16]. Although proposed in the 1960s, this approach was popularised through the NIST-funded Text REtrieval Conference (TREC) series of large-scale evaluation campaigns, which began in 1992 and has stimulated significant developments in IR over the past 20 years or so [17].

* Corresponding author.
E-mail addresses: kelbedweihy@cu.edu.eg (K.M. Elbedweihy), s.wrigley@sheffield.ac.uk (S.N. Wrigley), p.d.clough@sheffield.ac.uk (P. Clough), f.ciravegna@sheffield.ac.uk (F. Ciravegna).

[1] http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

http://dx.doi.org/10.1016/j.websem.2014.10.001
1570-8268/

However, despite the many benefits that come from the organisation of evaluation activities like TREC, the semantic search community still lacks a similar initiative on this scale. Indeed, Halpin et al. [18] note that "*the lack of standardised evaluation has become a serious bottleneck to further progress in this field*". In recent years evaluation activities have been organised to address this issue, including the SemSearch Challenge [18]; the SEALS semantic search evaluations [19,20]; the QALD open challenge [21] and the TREC Entity List Completion task [22,23]. However, these initiatives are yet to experience the level of participation shown by evaluation exercises in other fields. Furthermore, much of the experience gained from these initiatives is not accessible to the research community in general since they usually focus on reporting objectives and results with little explanation on the specific details of the evaluations, such as the methods and measures adopted. It is important to emphasise the need for more work towards evaluation frameworks/platforms that enable persistent storage of datasets and results that guarantee their reusability and the repeatability of tests. Forming agreement on the approaches and measures to be used within the search community for evaluating similar categories of tools and approaches is a key aspect of developing standardised evaluation frameworks [24]. In addition to the resources created, the value of organised evaluation campaigns in bringing together members of the research community to tackle problems collectively is also a stimulus for growth in a research field.

Although the focus of this paper is the evaluation of semantic search, we believe that there is much to learn from the wider IR community more generally. Therefore, this paper summarises IR evaluation activities and considers how this knowledge can be utilised in meeting the specific requirements of semantic search. The overall goal of this paper is to motivate the future development of a more formalised and comprehensive evaluation framework for semantic search. The remainder of this paper is structured as follows. An overview of evaluation in IR is provided in Section 2, followed by a discussion of important aspects of system-oriented evaluation, such as test collections and measures, in Section 3. Next, Section 4 describes approaches for user-oriented evaluation, such as the experimental setup and criteria to be assessed. Section 5 then goes on to summarise existing semantic search evaluation initiatives with potential limitations in existing approaches to evaluating semantic search and future directions discussed in Sections 6 and 7.

2. Approaches to IR evaluation

Evaluation is the process of assessing the 'worth' of something and evaluating the performance of an IR system is an important part of developing an effective, efficient and useable search engine [25,13]. For example, it is necessary to establish to what extent the system being developed meets the needs of its end users, quantify the effects of changing the underlying search system or its functionality, and enable the comparison between different systems and search strategies. How to conduct IR system evaluation has been an active area of research for the past 50 years or so, and the subject of much discussion and debate [25,13,15]. This is due, in part, to the need to incorporate users and user interaction into evaluation studies and the relationship between the results of laboratory-based vs. operational tests [26].

Harman [15] describes IR evaluation as "*the systematic determination of merit of something using criteria against a set of standards*". This implies the need for a *systematic approach* for conducting an evaluation, the need for suitable *criteria* for evaluating search and the need to evaluate with respect to *standards*, for example using a standard benchmark or comparing against a baseline system or approach. Cleverdon [27] identified six evaluation criteria that could be used to evaluate IR systems: (1) coverage, (2) time lag, (3) recall, (4) precision, (5) presentation, and (6) user effort. Of these, precision and recall have been the most widely used to evaluate IR systems. However, the success of an IR system, especially from a user's perspective, goes beyond the performance of indexing and retrieval and may include how well the IR system supports users in carrying out their search tasks and whether users are satisfied with the results [28,29].

Evaluation of search systems can be carried out at various levels and may involve multiple methods of evaluation in an iterative manner during development and subsequent deployment. Saracevic [25] distinguishes six levels of evaluation for information systems (including IR systems) as follows:

1. The *engineering level* deals with aspects of technology, such as computer hardware and networks to assess issues, such as reliability, errors, failures and faults.
2. The *input level* deals with assessing the inputs and contents of the system to evaluate aspects, such as coverage of the document collection.
3. The *processing level* deals with how the inputs are processed to assess aspects, such as the performance of algorithms for indexing and retrieval.
4. The *output level* deals with interactions with the system and output(s) obtained to assess aspects such as search interactions, feedback and outputs. This could include assessing usability.
5. The *use and user level* assesses how well the IR system supports people with their searching tasks in the wider context of information seeking behaviour (e.g., the user's specific seeking and work tasks). This could include assessing the quality of the information returned from the IR system for work tasks.
6. The *social level* deals with issues of impact on the environment (e.g., within an organisation) and could include assessing aspects such as productivity, effects on decision-making and socio-cognitive relevance.

Traditionally in IR evaluation there has been a strong emphasis on measuring system performance (levels 1–3), especially retrieval efficiency and effectiveness [13,15]. The creation of standardised benchmarks for quantifying retrieval effectiveness (commonly known as *test* or *reference collections*) is highly beneficial when assessing system performance [13,14]. However, evaluation at levels 4–6 is also important as it assesses the performance of the system from the user's perspective and may also take into account the user's interactions with the system, along with broader effects, such as its impact and use in operation [28,30,31]. In the following sections we discuss in more detail the two main approaches referenced in the literature: system-oriented and user-oriented evaluation.

3. System-oriented evaluation

System-oriented evaluation of IR systems has typically focused on assessing ranked lists of results given a specification of a user's query, although attention has also been given to evaluating IR systems that comprise of multiple finding aids, such as visualisations or facets, and for tasks beyond search, such as exploration and browsing [32]. One of the first and most influential proposals for system-oriented evaluation was based upon the Cranfield methodology [33]. The Cranfield approach to IR evaluation uses test (or reference) collections: re-useable and standardised resources that can be used to evaluate IR systems with respect to the system [16]. Over the years the creation of a standard test environment has proven invaluable for the design and evaluation of practical retrieval systems by enabling researchers to assess, in an objective and systematic way, the ability of retrieval systems to locate documents relevant to a specific user need.