Assisted Query Formulation Using Normalised Word Vector and Dynamic Ontological Filtering

Heinz Dreher and Robert Williams

Curtin University of Technology, GPO Box U1987
Perth, Western Australia 6845
{Heinz.Dreher, Bob.Williams}@cbs.curtin.edu.au

Abstract. Information seekers using the usual search techniques and engines are delighted by the sheer power of the technology at their command – speed, quantity. Upon closer inspection of the results, and reflection upon the next stages of the information seeking knowledge work, users are typically overwhelmed, and frustrated. We propose a partial solution by focusing on the query formulation aspect of the information seeking problem. First we introduce our version of a semantic analysis algorithm, named Normalised Word Vector, and explain its application in assisted query formulation. Secondly we introduce our ideas of supporting query refinement via Dynamic Ontological Filtering.

1 Introduction

Information and Communications Technologies (ICT) pervade our society in which a growing proportion of the work is mental work (typically referred to as knowledge work) as opposed to physical work. The record of empowerment of physical workers through technology clearly shows the enormous benefit in terms of increased production, efficient and effective utilisation of resources, and greater safety for workers. Knowledge workers should expect to see similar gains in capacity, productivity, and in the quality of outputs, but progress in the empowerment of the knowledge worker needs a boost so that this promise and expectation can be realised. It is not so much a case of lack of vision:

> Neither the naked hand nor the understanding left to itself can effect much. It is by instruments and helps that the work is done, which are as much wanted for the understanding as for the hand. And as the instruments for the hand either give motion or guide it, so the instruments of the mind supply either suggestions for the understanding or cautions. [1].

More recently, by some three and a quarter centuries, Vannevar Bush [2] shared with us his vision of how research workers could be empowered with his MEMEX device. But despite the vision, and the tremendous advances in ICT, and their now pervasive nature, the knowledge worker is left languishing by and large. For example, whilst a literature search can now be conducted in a matter of days if not hours where only two decades ago it took weeks if not months, the researcher is confronted with millions upon millions of 'hits'.

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 282 – 294, 2006.

Present methods of refinement of the result of a query or search are inadequate – there is far too much material which is potentially relevant. Additionally, users, don't really know what is relevant until some way through the discovery, learning, or knowledge acquisition process. Help is needed with the Query-formulation → Find → Re-formulation phases of knowledge work.

In this article we consider the query-formulation aspect of the general problem. There are two contributions we propose to integrate into search and find processes. Firstly, an adaptation of a semantic analysis algorithm named Normalised Word Vector (NWV) developed by Williams [7] for the MarkIT (www.essaygrading.com) Automated Essay Grading project [8] with the aim of accepting natural language query expressions. Secondly, we propose a dynamic ontological filter to be applied to the query in order to maximise search relevance. Categorizing the results returned by search engines and presenting the categories to the user through a special browser endowed with an ontology navigation scheme is expected to contribute to a refinement of the search query and hence improve relevance and thus information quality as required by the user.

2 Assisted Query Formulation – Empowering P to Refine Q

Imagine a human user P is interested in researching Yoga, and enters this as a search-term (Q) into Google which delivers circa 39 million results in one tenth of a second. This is impressive until P begins to use the returned results. P quickly determines that some strategy is needed to reduce the quantity and increase the relevance of the results, but to accomplish this P will need to refine Q, and proceeds using the traditional methods such as including Boolean operations, appending adjectives, enclosing search strings within quotes, and so on. P's focus has now been redirected from the concept "Yoga", as originally envisaged, to the science of search-term and query formulation. Actually, we all know that P did not have just "yoga' in mind, but some idiosyncratic aspect/s thereof. Surely we can better support P in the information seeking task pertaining to Yoga.

Consider the 7-Step process in Table 1, which we have termed Query-formulation → Find → Re-formulation (QFR).

Continuing our Yoga example from above, at Step 1) we have P composing a natural language query into the special browser.

Q = I would like to take an English language based Yoga-teacher course as soon as possible

(Google returns 900 odd results in about half a second, and if Q is enclosed in quotes, Google returns no matching documents, also quite speedily).

The NWV technology computes Normalised Concepts based on Q, which for example may be Language, Lifestyle, Time, Education. These constitute the 'core concepts' contained in Q - comprising Step 2). Varied forms of Q as expressed by various P would all yield the same Normalised Concepts thereby already greatly simplifying and focusing the subsequent document match and retrieval. At this point one may mention that the 'core concepts' could be readily translated into arbitrarily many natural languages, and depending on the respective thesaurus (or alternative) structures, similarly high quality results could be expected for those alternative