Advanced Engineering Informatics 30 (2016) 737–750

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

Full length article

Enhanced context-based document relevance assessment and ranking for improved information retrieval to support environmental decision making

Xuan Lv, Nora M. El-Gohary *

Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, 205 N. Mathews Ave., Urbana, IL 61801, United States

ARTICLE INFO

Article history:
Received 3 January 2016
Received in revised form 25 August 2016
Accepted 25 August 2016

Keywords:
Information retrieval
Context-based relevance assessment
Context-enhanced document ranking
Vector space model
Statistical language model
Project environmental review

ABSTRACT

There is a need for enhanced context-based document relevance assessment and ranking to facilitate the retrieval of more relevant information for supporting environmental decision making. This paper proposes a new context-based relevance assessment method, which allows for enhanced context representation and context-based document relevance recognition through: (1) a context-aware and deep semantic concept indexing approach, and (2) a deep and semantically-sensitive relevance estimation approach. The proposed relevance assessment method was integrated into two widely-used document ranking models [vector space model (VSM) and statistical language model (SLM)], resulting in two improved ranking methods: (1) a context-enhanced VSM-based method, and (2) a context-enhanced SLM-based method. The two context-enhanced document ranking methods were evaluated in retrieving webpages that are relevant to transportation project environmental review. The two context-enhanced methods were compared with each other and with their provenance methods (i.e., original VSM and SLM) in terms of mean precision (MP) and mean average precision (MAP). The context-enhanced VSM-based method outperformed the context-enhanced SLM-based method on every metric. It achieved 48% MAP, 79% MP at the top 10 retrieved documents, and over 65% MP at the top 50 retrieved documents, on the testing data. It also showed significant improvement over the state-of-the-art keyword-based VSM method.

1. Introduction

The National Environmental Policy Act (NEPA) requires every transportation project obtaining federal funding or requiring federal approval to undertake an environmental review process to evaluate alternatives and their potential impacts on the environment [1]. Although the environmental review process successfully brings environmental considerations into transportation decision making through involving a wide range of stakeholders from government agencies to the interested public, it has been criticized as "a cause of delays for projects because of time-consuming requirements" [2]. According to a number of studies on accelerating transportation project delivery [3–6], the environmental review process has been recognized as one of the major causes for the lengthy and costly project development process. According to a study conducted by the U.S. Government Accountability Office [7], the median time to complete the environmental review process for large-scale highway projects was over 7 years in 2013 and the process may cost several million dollars.

The environmental review process mandates intensive studying and clear documentation of the environmental impacts, which requires transportation practitioners from a variety of different agencies to collect, analyze, and communicate a massive amount of textual resources such as environmental laws and regulations, best management practices and guidance documents, and project documents. The key to improve the environmental review process is to provide transportation practitioners with the right information at the appropriate decision points [8]. Duplication of efforts can be avoided by learning from previous cases, i.e., environmental reviews conducted for similar types of projects that potentially impact similar environmental resources. This requires searching for and finding such relevant environmental reviews and associated documents. However, substantial gaps exist in the ability of transportation practitioners to reliably and efficiently find the right information to support such mission-critical analysis [9]. As indicated by a National Cooperative Highway Research Program

* Corresponding author.
E-mail addresses: xuanlv2@illinois.edu (X. Lv), gohary@illinois.edu (N.M. El-Gohary).

http://dx.doi.org/10.1016/j.aei.2016.08.004
1474-0346/

(NCHRP) study [10], 80–90% of the information that transportation practitioners have access to is unstructured, and they may spend up to 35% of their time searching for the right information to support decision making. For the project environmental review process, specifically, the problem is further compounded by the large amount of information that is involved and the complex nature of the process.

There is, thus, a need for information retrieval (IR) methods that better understand the knowledge of the environmental review process and the context of its retrieval and use by transportation professionals, in order to improve the efficiency and effectiveness of finding information for supporting environmental decision making [11]. However, most of the existing IR efforts are limited in context representation and context-based relevance assessment. To address this gap, this paper: (1) proposes a context-based relevance assessment method to improve concept representation and document relevance recognition for supporting context-enhanced document ranking in the transportation environmental review domain, (2) integrates the proposed context-based relevance assessment method into two widely-used document ranking models [vector space model (VSM) and statistical language model (SLM)] to facilitate context-enhanced semantic document ranking, and (3) compares both enhanced ranking methods (context-enhanced VSM-based and SLM-based methods), to each other and to their provenance methods (i.e., original VSM and SLM), in terms of their IR performance for comparative evaluation.

2. Background

2.1. Information retrieval

IR is the activity of obtaining information sources (usually documents) of an unstructured nature (usually text) that is relevant to the user's information need from within large collections [12]. The user's information needs are represented as queries, and the IR system presents the documents to the user according to their relevance to the queries. Current IR systems rely largely on keyword-based document ranking methods, which evaluate the textual relevance of documents to the users' queries based on keywords, and can misunderstand a user's true intent due to their limited capabilities for incorporating content semantics and contextual information [13–15]. Due to this weakness, current IR systems can be very ineffective when dealing with context-sensitive searching tasks, such as searching for relevant documents to specific domains. Many research efforts have, thus, been conducted to improve the keyword-based document ranking methods or develop semantic-based document ranking methods in order to more precisely evaluate the relevance based on content semantics and contextual information [16–26].

For the transportation environmental review domain, as indicated by recent studies [11,27], the ineffectiveness of current IR systems are aggravated when searching for relevant information to support decision making for the domain. For example, the following use case scenario provides an illustrative example: an environmental specialist (user's role) from the Illinois Department of Transportation (IDOT) is working on a new toll road corridor (project type) that affects nearby wetlands (affected resource) in northeastern Illinois (project location), he/she would like to find similar projects that also affect wetlands and how their environmental impacts are evaluated, and he/she searchers Google for "highway projects have environmental impact on wetlands". Fig. 1 shows the first result page that was retrieved by Google. All the retrieved results in the first page only provide general information about evaluating environmental impacts on wetlands, such as guidance on quantifying the impacts on wetland loss (first and fifth results), and mitigation measures for the impacts on wetlands (second result); and none of them provide the specific project examples that the environmental specialist needs to retrieve. To improve the retrieval results, he/she enhances the query and searches Google for "Illinois tollway projects have environmental impact on wetlands". Fig. 2 shows the first result page using the enhanced query. Although the third and sixth retrieved results provide information on the specific projects the environmental specialist is looking for, other results only provide general information such as guidance on wetland restoration (first and second results) and the environmental studies manual (fifth). The highly context-sensitive nature of the transportation environmental review process and of the searching process of related information makes it difficult to retrieve satisfactory results using conventional IR systems. The searching process of the environmental review relevant information is sensitive to the context of the domain knowledge (e.g., project type, project location, environmental review type, affected resources, etc.), the context of the user (e.g., user role, user task at hand, user profile), and the context of the searching process (e.g., searching location, searching environment, searching device). For example, in the above use case scenario, the information on the desired highway projects is sensitive to the project type, project location, environmental resources affected, and user role. An enhanced semantic-based document ranking method is needed to help retrieve more relevant results by adapting to these various contexts.

2.2. Document ranking models

A document ranking model provides the basic notion of what it means for a document to be relevant to a query. Among the many different document ranking models proposed in the literature, the VSM and the SLM are the most studied and widely used. The VSM is a similarity-based model that assumes that the relevance of a document to a query is correlated with the similarity between the query and the document at some level of representation [28]. In the VSM, a document and a query are represented as two vectors of terms, which are typical words and phrases. Each term is assigned a weight that reflects its "importance" to the document or the query. This model measures the relevance of a document to a query as the similarity between the query vector and the document vector. The cosine similarity and the inner-product between the two vectors are often used as the similarity measures [29].

The SLM is a probabilistic model that assumes that the documents in a collection should be ranked by the decreasing probabilities of their relevance to a query [30]. A document is generally viewed as a sample from a language model, which estimates the distribution of words in a given language. Based on this assumption, this model measures the relevance of a document to a query as the likelihood that the query was generated based on the estimated language model of each document [31].

The VSM and the SLM each offers different advantages in different situations. Previous studies [19,32,33] indicated that there is no single model that outperforms the other in all applications. For example, Raghavan and Iyer [33] found that the VSM had better performance when retrieving relevant advertisement for sponsored search; while Lin et al. [19] found that the SLM was better at retrieving passages of technical documents for architecture/engineering/construction (AEC) projects and research. Because the performances of the two models could vary from domain to domain and application to application, it is necessary to compare the performances of the two models in facilitating context-enhanced semantic document ranking in the transportation project environmental review (TPER) domain.