Computers in Industry 78 (2016) 29–42

Contents lists available at ScienceDirect

Computers in Industry

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

Integrating a semantic-based retrieval agent into case-based reasoning systems: A case study of an online bookstore

Jia Wei Chang [a], Ming Che Lee [b,*], Tzone I Wang [a]

[a] Department of Engineering Science, National Cheng Kung University, Tainan, Taiwan
[b] Department of Computer and Communication Engineering, Ming Chuan University, Taoyuan, Taiwan

ARTICLE INFO

Article history:
Received 27 November 2014
Received in revised form 4 October 2015
Accepted 13 October 2015
Available online 26 October 2015

Keywords:
Short-text semantic similarity
Case-based reasoning
Case-based machine learning

ABSTRACT

Natural language search engines should be developed to provide a friendly environment for business-to-consumer e-commerce that reduce the fatigue customers experience and help them decide what to buy. To support product information retrieval and reuse, this paper presents a novel framework for a case-based reasoning system that includes a collaborative filtering mechanism and a semantic-based case retrieval agent. Furthermore, the case retrieval agent integrates short-text semantic similarity (STSS) and recognizing textual entailment (RTE). The proposed approach was evaluated using competitive methods in the performance of STSS and RTE, and according to the results, the proposed approach outperforms most previously described approaches. Finally, the effectiveness of the proposed approach was investigated using a case study of an online bookstore, and according to the results of case study, the proposed approach outperforms a compared system using string similarity and an existing e-commerce system, Amazon.

1. Introduction

Business-to-consumer (B2C) refers to a business model where companies and consumers trade digitally, which also describes a company that provides goods and services for consumers on the internet, one of the best examples of B2C e-procurement is the online bookstore of Amazon.com [1]. The focus of B2C e-procurement is to suggest enticing prospects to customers and to retain them and the share values they created [2], the final goal of which is to convert shoppers into buyers actively and constantly. Related studies have demonstrated that the recommendation quality of a commerce system directly affects customer satisfaction of a website [3–5]. However, most existing commerce systems use keyword searches, which have performance limitations because it is not possible for users to ask their question by using natural language and to acquire answers instantly by logically matching potential words that might be related. With the development of natural language processing (NLP), natural language search is a possible solution to the problems, thus allowing users to express what they want in their own words. In addition, natural language search can "read the minds" of users, enabling them to use internet technology more easily, thus improving the recommendation quality, reducing the fatigue associated with search engine use, and transforming the search experience into an effective, positive, and more human experience.

Existing recommender systems are mainly classified into collaborative filtering (CF) techniques and content-based (CB) methodologies. CF techniques recommended items by the known preferences of users. However, new items are not included because they have to be rated by many users before they can be recommended. This is called the cold-start problem, which limits their performance [6]. By contrast, CB methods recommend new items because recommendations are based on the descriptive characteristics of items, which rely on more specific information about items. However, they must overcome the problems of limited diversity and possible overspecialization. Recent studies have demonstrated that a hybrid approach can combine the advantages of both techniques, overcome the limitations of CF and CB, and improve the accuracy of recommendation [7,8].

Case-based reasoning (CBR) is a popular model that develops commerce recommendation systems [9], and is a framework with a high compatibility of combining CF [6,10,11] as well as CB [9,12,13]. In this technique, new problems are solved by utilizing or modifying the solutions of similar existing problems, the core of which is using similarity measure to quantify the differences that exist between objects [14] because CBR uses similarity measures to

* Corresponding author.
E-mail address: leemc@mail.mcu.edu.tw (M.C. Lee).

http://dx.doi.org/10.1016/j.compind.2015.10.007
0166-3615/

identify cases that are similar to the problem at hand, and most of measures that evaluate similarities of non-numeric properties are syntactic methods (such as brute force, longest common subsequence, and Levenshtein distance) and calculate the string similarity between two words [15–17] but often fail to match exactly when confronted with the words that have associated meanings. To overcome the limitations of syntactic approaches, semantic similarity measure can be used that is more suitable for describing non-numeric properties than syntactic similarities, which also can help system to communicate directly with users through natural languages and to eliminate the burden of query formulation and the enormous document load needed to find the answers. The NLP-based technologies facilitate the CBR model that can streamline the integration and creation of knowledge-based systems [18]. Recent business systems have evolved out of NLP techniques that demonstrated the potential in business workflow [19–23].

This study proposes a hybrid CBR model that combines the CB and CF mechanisms to accept natural language queries and increase the recommendation accuracy. Mainly, we designed a semantic similarity algorithm to analyze the content of user queries and goods, which plays the role of a case retrieval agent in the proposed system. The algorithm can more accurately understand users' intention to retrieve the most appropriate cases for reuse, which facilitates our CBR system in that it does not require a high-quality case base in the beginning, thus solving the cold-start problem of the CF technique. Furthermore, the proposed system includes the CF mechanism for recommendation ranking. This not only avoids the problems of CB, such as limited diversity and overspecialization, but also makes our system stronger. Finally, the effectiveness of the proposed system is illustrated using a case study of an online bookstore service, such as Amazon, but it can be developed for any domain. This study aimed to (1) use semantic similarity measurements instead of string similarity measurements to retrieve cases that can facilitate our system in having a more satisfactory recommendation at the beginning; (2) integrate CB and CF mechanisms into a hybrid CBR system to constantly improve the recommendation quality; and (3) propose a NLP-based CBR approach that can accept natural language queries to achieve user friendliness.

The paper is organized as follows. Section 2 describes the theoretical framework and related works of the CBR system, word-sense disambiguation, and short-text semantic similarity. Section 3 provides the details of the development of the proposed system and retrieval method. Section 4 describes the performance test of the proposed STSS algorithm by using established benchmarks. Section 5 describes the evaluation of the semantic measures proposed in this paper and the effectiveness of the proposed approach is illustrated using a case study on an online bookstore. Finally, Section 6 presents the conclusions.

2. Backgrounds

This section provides a focused introduction to the relevant foundations of CBR systems and word-sense disambiguation, as well as a review of short-text semantic similarity.

2.1. Case-based reasoning systems

For application-oriented projects, most systems are based on a CBR architecture [24–28], which indicates that CBR is a well-established model that facilitates methodology design in industry engineering and B2C commerce. CBR is a branch of artificial intelligence (AI), which is a method based on using past experience for problem solving and decision making, searching for the solutions to novel problems using previously solved problems, and reusing the existing solutions in new situations [29,30]. The outstanding characteristic of CBR is that it does not need to match the user's query exactly, such as in searching problems. These cases are usually similar to some extent. Because the basic assumption of CBR is that similar problems have similar solutions, even if the repository does not contain a solution that immediately addresses a user's problem, a similar answer can be available for use as a starting point. Similar solutions can then be adapted and at least provide some inspiration and guidance for the user. A CBR system usually consists of domain/expert knowledge, a case-base of past experiences, and a similarity measure for searching related cases [31]. Domain knowledge refers to knowledge about the features of the different entities and what is a "case". A case-base contains a set of cases, each of which describes a problem, a solution to the problem, and annotations about how the solution was derived. Similarity measures are developed according to the features of the case because the problem is typically defined in terms of specific features of objects, and those features can be numeric or non-numeric properties; the similarity measure is used to identify the cases that are most relevant to the problem. The CBR system responds to the query using a given algorithm and a similarity measure, which consists of translating and matching a query against a set of information objects. Finally, a similarity measure calculates the similarities that exist between objects. To realize automated reasoning, CBR systems basically have been formalized as a four-step process (Fig. 1) [32]:

Retrieve: Given a specific problem, retrieve similar cases from the case-base/repository to solve the problem.
Reuse: Choose the possible solutions from the retrieved cases. If the solutions are not able to be used directly, they need to be adapted for the new situation.
Revise: Modify the existing solution to the target problem, test the new solution for the problem, and if necessary, continue to revise.
Retain: Store the resulting new cases in the repository if the solution has been successfully applied to the target problem.

However, ambiguity often occurs in query design when users have no idea about the exact expressions and the related concepts they want to know, but they may have some contextual clues, such as the function of the objective. Therefore, traditional CBR uses syntactic similarity measurements, which are not intelligent enough for meaning-related searches. In contrast, because engineering designers usually need to construct a case-base for the CBR system and the task often needs many manual works and expert experiences, automatic knowledge acquisition has become an emerging issue for the development of CBR systems. Moreover, solutions from past cases may not directly be reusable; in these situations, they should be adapted to better fit the new problem.

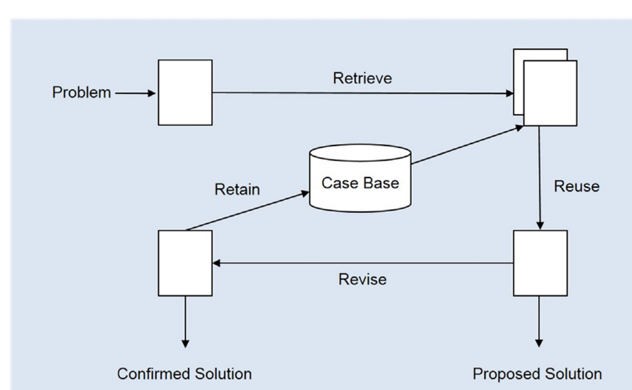

Fig. 1. The traditional CBR mechanism.