Similarity-based approach for accurately retrieving similar cases to intelligently handle online complaints

Xiaolan Cui [a]**, Shuqin Cai** [a,*]**, Yuchu Qin** [b]

[a] *School of Management, Huazhong University of Science and Technology, Wuhan 430074, PR China*

[b] *School of Computing and Engineering, University of Huddersfield, Huddersfield HD1 3DH, UK*

Abstract

Purpose: The purpose of this paper is to propose a similarity-based approach to accurately retrieve reference solutions for the intelligent handling of online complaints.

Design/methodology/approach: This approach uses a case-based reasoning framework and firstly formalizes existing online complaints and their solutions, new online complaints, and complaint products, problems, and content as source cases, target cases, and distinctive features of each case, respectively. Then the process of using existing word-level, sense-level, and text-level measures to assess the similarities between complaint products, problems, and contents is explained. Based on these similarities, a measure with high accuracy for assessing the overall similarity between cases is designed. The effectiveness of the approach is evaluated by numerical and empirical experiments.

Findings: The evaluation results show that a measure simultaneously considering the features of similarity at word, sense, and text levels can obtain higher accuracy than those measures considering only one level feature of similarity and the designed measure is more accurate than all of its linear combinations.

Practical implications: The approach offers a feasible way to reduce manual intervention in online complaint handling. Complaint products, problems and content should be synthetically considered when handling an online complaint. The designed procedure of the measure with high accuracy can be applied in other applications that consider multiple similarity features or linguistic levels.

Originality/value: A method for linearly combining the similarities at all linguistic levels to accurately assess the overall similarities between online complaint cases is presented. This method is experimentally verified to be helpful to improve the accuracy of online complaint case retrieval. This is the first consideration of the accuracy of the similarity measures for online complaint case retrieval.

Keywords: Online complaint handling; Case-based reasoning; Case retrieval; Similarity measure; Accuracy

Paper type: Research paper

1. Introduction

The development of network technology provides new avenues for customers to freely exchange unfavorable experiences about the products or services of enterprises. Complaints have been changed from private communications offline only with family members and friends into public shares via network platforms such as websites, forums, blogs, and content communities. Such publicly shared online complaints, also known as electronic negative word-of-mouth (Jansen et al., 2009), are found to have detrimental effect on the decision procedure of customers and confirmed to cause costly irretrievable loss to enterprises (Cheung and Thadani, 2012; Van Noort and Willemsen, 2012). Hence, it is of great necessity for enterprises to handle online complaints.

---

* Corresponding author. Tel.: +86 027 87556481

E-mail address: caishuqin@hust.edu.cn (S. Cai).

In recent years, enterprises have begun to actively monitor and take interventions to handle online complaints (Fournier and Avery, 2011). Thomas et al. (2012) summarized five strategies for enterprises' interventions that include (1) delaying response, (2) listening, acknowledging, and addressing negative feedback, (3) associating with customers, (4) taking legal action, and (5) removing or suppressing unwanted information. Enterprises can choose right strategies to deal with online complaints according to specific situations. Even though these five strategies are capable of solving the problems that are involved in online complaints to some extent, they are all one-to-one strategies that heavily depend on the professional skills and manual labor of employees. It is very difficult for the employees of an enterprise to use the five strategies to manually handle massive amounts of online complaints about the products or services of this enterprise in time.

Aiming at dealing with online complaints intelligently and timely, a number of approaches (Ma et al., 2006; Coussement and Van den Poel, 2008; Sultan et al., 2008; Zirtiloğlu and Yolum, 2008; Galitsky et al., 2009; Minaei-Bidgoli and Akhondzadeh, 2010; Sitko-Lutek et al., 2010; Trappey et al., 2010; Chen et al., 2012; Lee et al., 2015; Faed et al., 2016) have been proposed during the past decade. These approaches generally used some computer technologies, such as machine learning (Coussement and Van den Poel, 2008; Galitsky et al., 2009), similarity assessment (Galitsky et al., 2009), data mining (Minaei-Bidgoli and Akhondzadeh, 2010; Trappey et al., 2010; Chen et al., 2012), social network analysis (Sitko-Lutek et al., 2010), principal component and data envelopment analysis (Faed et al., 2016), intelligent agent (Sultan et al., 2008), ontology (Ma et al., 2006; Zirtiloğlu and Yolum, 2008; Lee et al., 2015), and case-based reasoning (CBR) (Lee et al., 2015), to model existing online complaints and their solutions and retrieve reference solutions for a new online complaint from these models to assist employees to handle online complaints. Each of the approaches can work well in its specific context and can reduce the dependence on the professional skills and manual labor of employees to some extent. However, the retrievals of reference solutions in them are mostly based on keywords, which is difficult to obtain desired retrieval accuracy in most cases (Zhang et al., 2013). An accurate retrieval should not only consider keywords, but consider the linguistic elements of word, sense, and text simultaneously (Etzioni, 2011).

In response to the above limitation, a similarity-based approach for retrieving reference solutions accurately to deal with online complaints intelligently is proposed in this paper. Compared with the existing approaches (Ma et al., 2006; Coussement and Van den Poel, 2008; Sultan et al., 2008; Zirtiloğlu and Yolum, 2008; Galitsky et al., 2009; Minaei-Bidgoli and Akhondzadeh, 2010; Sitko-Lutek et al., 2010; Trappey et al., 2010; Chen et al., 2012; Lee et al., 2015; Faed et al., 2016), the proposed approach synthetically considers the similarities at all linguistic levels (i.e. the similarities of words, senses, and texts). It experimentally verifies that an approach taking into account the similarities at all linguistic levels can obtain higher accuracy than those approaches merely considering the similarity of keywords. The approach also presents a mechanism to linearly combine the similarities at all linguistic levels to accurately assess the overall similarity between online complaint cases. This mechanism is also experimentally verified to be helpful to improve the accuracy of online complaint case retrieval.

The rest of the paper is organized as follows. An overview of related work is provided in Section 2. The details of the proposed approach are explained in Section 3. Section 4 reports a prototype implementation of the approach and evaluates the approach by experiments. Section 5 ends the paper with a conclusion.

2. Related work

Online complaint handling is an important part in the customer relationship management of most enterprises nowadays. During the past two decades, a number of approaches have emerged in this area. These approaches can be classified into manual and intelligent handling approaches based on the degree of employees' participation.

Manual handling approaches are the most commonly used approaches for online complaint handling in practice. They attempt to deal with online complaints through the participation of first line employees. In this context, the professional skills of first line employees are the key element to decide whether an online complaint handling is successful. Many practical strategies, such as taking webcare (Fournier and Avery, 2011; Van Noort and Willemsen, 2012), using legal weapon and partnering with users (Thomas et al., 2012), monitoring actively and responding initiatively (Harrison-Walker, 2001; Lee and Lee, 2006), responding in an accurate and kind manner and leveraging online chat service (Cho et al., 2002; Cheng and Loi, 2014), responding specifically (Mattila et al., 2013; Chang et al., 2015), and launching long-term mechanisms and increasing interactions with complainants (Pfeffer et al., 2014; Schwab, 2015), have been presented to improve the professional skills of first line employees. It can be reflected from the listed strategies that manual handling approaches have good flexibility. In such approaches, the most suitable solutions can be used by the well-trained employees to deal with online complaints. However, manual handling approaches heavily rely on the manual labor of employees. It is difficult for employees to manually handle explosively growing online complaints in time.

Intelligent handling approaches aim to overcome the drawback of manual handling approaches through leveraging some computer technologies, such as machine learning (Coussement and Van den Poel, 2008; Galitsky et al., 2009), similarity assessment (Galitsky et al., 2009), data mining (Trappey et al., 2010; Minaei-Bidgoli and Akhondzadeh, 2010; Chen et al., 2012), social network analysis (Sitko-Lutek et al., 2010), principal component and data envelopment analysis (Faed et al., 2016), intelligent agent (Sultan et al., 2008), ontology (Ma et al., 2006; Zirtiloğlu and Yolum, 2008; Lee et al., 2015), and case-based reasoning (CBR) (Lee et al., 2015), to model historical online complaints and their solutions and retrieve reference solutions for a new online complaint from these models to assist online complaint handling. The major advantage and common limitation of these approaches were analyzed and highlighted in the introduction.

The present paper continues this line of research and attempts to overcome the common limitation of the existing intelligent handling approaches. Its main contribution can be briefly summarized as: The paper presents a method to linearly combine the similarities at all linguistic levels to accurately assess the overall similarities of online complaint cases for retrieving similar cases to assist online complaint handling. The retrievals of reference solutions in the existing approaches are mostly based on keywords, which could result in low retrieval accuracy (Zhang et al., 2013). Differently, the presented method simultaneously takes into account the similarities of complaint products, problems, and contents, which include word-level, sense-level, and text-level similarities. It is experimentally verified to obtain higher accuracy than those methods taking into account keyword similarities only. Moreover, the method also explains how to establish an overall similarity measure for different case bases under the prerequisite of ensuring the highest accuracy among the accuracies

of all possible linear combinations of these measures. To the best of the knowledge, this is the first consideration of the accuracy of the similarity measures for online complaint case retrieval.

3. Similarity-based approach

This section describes a similarity-based approach for accurately retrieving similar cases to intelligently deal with online complaints. The schematic representation of this approach is shown in Figure 1. The first step is to construct an online complaint handling ontology and an online complaint corpus which are needed in the similarity-based approach. The similarities between complaint products, between complaint problems, and between complaint contents are respectively assessed in the second, third, and fourth steps. Then the overall similarity between online complaint cases is defined as a weighted sum of these three similarities. The last step is to find out a measure with high accuracy to assess the overall similarity between online complaint cases. Using the found out measure, similar source cases for each target case, which are constructed through leveraging Gupta's extraction method (Gupta, 2013) to extract the complaint products and problems from each new online complaint, can be accurately retrieved. These similar source cases can be used to process the problems in the new target case by a case reuse or a case revision mechanism. The new target case and its reference solutions can then be retained as a new source case through a case retention mechanism. Because this paper mainly aims to explain how to accurately retrieve similar source cases for the intelligent handling of online complaints, the processes of case reuse, revision, and retention will not be discussed.

The details of the five steps are explained in the following sub-sections: (1) the construction of an online complaint handling ontology and an online complaint corpus; (2) the assessments of the similarities between products and between problems; (3) the assessment of the similarity between contents; and (4) the assessment of the overall similarity between cases.

3.1. Ontology and corpus

3.1.1 Online complaint handling ontology

Ontology, an explicit specification of a conceptualization of terms and their relationships, is well-known for having rigorous computer-understandable semantics (McGuinness and van Harmelen, 2004). Even though the application of ontology has its root in the field of the Semantic Web, it has been extended to many other fields during the past decade. In the area of customer complaint handling, some researchers like Ma et al. (2006), Zirtiloğlu and Yolum (2008), and Lee et al. (2015) have constructed their respective complaint handling ontologies to explicitly express the semantics of customer complaints to facilitate the intelligent handling of such complaints. Each ontology can work well in its specific context. But these ontologies cannot be directly used in the proposed similarity-based approach because the roles of ontology in this approach are different from that in (Ma et al., 2006; Zirtiloğlu and Yolum, 2008; Lee et al., 2015).

In the similarity-based approach, an online complaint handling ontology was manually constructed and mainly plays three roles of case base of CBR, semantic network of complaint products, and semantic network of complaint problems:

Role 1. *Case base of CBR*. The purpose of using ontology here is to explicitly represent the semantics of online complaint cases. For example, an online complaint case is defined as a seven-tuple $\mathbf{C} = (N, D, C, \mathbf{D}, \mathbf{P},$

T, $\mathbf{S}$) in the case base, where N is the case number, D is the last revised date of the case, C is the complainant of the complaint in the case, $\mathbf{D} = \{D_1, D_2,\ldots, D_m\}$ is a set of m complaint products in the case, $\mathbf{P} = \{P_1, P_2,\ldots, P_u\}$ is a set of u complaint problems in the case, T is the complaint content in the case, and $\mathbf{S} = \{S_1, S_2,\ldots, S_x\}$ is a set of x solutions of the complaint problems in the case. The semantics of such definition can be explicitly represented using the web ontology language (OWL) (McGuinness and van Harmelen, 2004) and the concepts and relations in the representation can be instantiated by concrete online complaints and their solutions. In the constructed case base of CBR, the concepts and relations were instantiated by 500 online complaints about the products provided by AT&T that were collected from Better Business Bureau and the solutions of the problems in these complaints that were manually collected from the Internet. After the instantiation, the case base was encoded in OWL format, a computer-readable and computer-understandable format, which lays a solid foundation for the intelligent retrieval, reuse, revision, and retention of cases in the case base.

A similarity measure with high accuracy for assessing the overall similarity between online complaint cases

K candidate measures (e.g. $Sim_{Sanchez}$, Sim_{Lin})

L candidate measures (e.g. $Sim_{Sanchez}$, Sim_{Lin})

M candidate measures (e.g. Sim_{Islam}, $Sim_{Mihalcea}$)

N sample case pairs Judged similarities

Similarity between complaint products

Similarity between complaint problems

Similarity between complaint contents

$K \times L \times M$ groups of weights

Online complaint handling ontology (Case base; Semantic networks)

WordNet

Corpus

Canonical correlation analysis method

Source cases

Target case

Case retention

Similar source cases

Feature extraction

Employees

Solutions can be directly reused?

Complainant

NO

YES

Case revision

Case reuse

Reference solutions

Figure 1. The schematic representation of the similarity-based approach.

Role 2. *Semantic network of complaint products*. Semantic network is a network that represents the semantic relationships between different concepts. It usually exists in the form of a directed or undirected graph consisting of vertices and edges, which represent concepts and semantic relationships between concepts, respectively. Representative examples of semantic network are WordNet (Fellbaum, 1998) and HowNet (Dong and Dong, 2006), which are respectively established to represent the semantic relationships between the different general concepts in the English and Chinese languages. Even though WordNet has been, as shown in Figure 1, used in the assessment of the similarity between complaint contents, it is still necessary to construct a specialized semantic network of complaint products for assessing the similarity between them since WordNet might not contain all concepts about complaint products. In the similarity-based approach, a semantic network of the complaint products provided by AT&T was manually established through learning all the products of AT&T and their semantic relationships. This network was also encoded in OWL format.

Role 3. *Semantic network of complaint problems*. A semantic network of complaint problems is built for assessing the similarity between them. In the similarity-based approach, a semantic network of the complaint problems about the products provided by AT&T was manually constructed by learning all the complaint problems in the collected 500 online complaints. This network was also encoded in OWL format.

3.1.2 Online complaint corpus

In natural language processing, a corpus is a large, structured, and electronically stored and processed set of texts. It is usually used in the occurrence checking, hypothesis testing, and statistical analysis of language elements (e.g. words, sentences). Most of the corpus-based similarity measures work on the basis of the occurrence checking of language elements in their specific corpuses. In the proposed similarity-based approach, an online complaint corpus is, as shown in Figure 1, needed to be established to form the basis for assessing the similarity between complaint contents.

The similarity-based approach established the online complaint corpus for dealing with the online complaints about the products of AT&T through fetching 5000 online complaints about the products provided by AT&T from Better Business Bureau and storing each online complaint as one separate line in a text document.

3.2. Similarities between products and between problems

An online complaint case may contain one or more complaint products (problems). Thus, the similarity between the complaint products (problems) in two online complaint cases is the similarity between the sets of the complaint products (problems) in these two cases. Formally, let $\mathbf{C}_1 = (N_1, D_1, C_1, \mathbf{D}_1, \mathbf{P}_1, T_1, \mathbf{S}_1)$, $\mathbf{C}_2 = (N_2, D_2, C_2, \mathbf{D}_2, \mathbf{P}_2, T_2, \mathbf{S}_2)$, $\mathbf{D}_1 = \{D_{1,1}, D_{1,2}, \ldots, D_{1,m}\}$, $\mathbf{D}_2 = \{D_{2,1}, D_{2,2}, \ldots, D_{2,n}\}$, $\mathbf{P}_1 = \{P_{1,1}, P_{1,2}, \ldots, P_{1,u}\}$, and $\mathbf{P}_2 = \{P_{2,1}, P_{2,2}, \ldots, P_{2,v}\}$. Then the similarity between the complaint products (problems) in $\mathbf{C}_1$ and $\mathbf{C}_2$ is the similarity between $\mathbf{D}_1$ and $\mathbf{D}_2$ ($\mathbf{P}_1$ and $\mathbf{P}_2$). Because each complaint product (problem) is a concept of the constructed online complaint handling ontology, the computation of the similarity between $\mathbf{D}_1$ and $\mathbf{D}_2$ ($\mathbf{P}_1$ and $\mathbf{P}_2$) can be transformed to the assessment of the similarity between ontology concepts. Without loss of generality, the computation of the similarity between $\mathbf{D}_1$ and $\mathbf{D}_2$ is taken as an example to illustrate the transformation process, which contains the following three steps:

Step 1. *Establish a similarity matrix of complaint products*. Because $\mathbf{D}_1 = \{D_{1,1}, D_{1,2}, \ldots, D_{1,m}\}$ and $\mathbf{D}_2 = \{D_{2,1}, D_{2,2}, \ldots, D_{2,n}\}$, a similarity matrix between $\mathbf{D}_1$ and $\mathbf{D}_2$ can be established as follow:

$$\mathbf{M_D} = \begin{bmatrix} Sim(D_{1,1}, D_{2,1}) & Sim(D_{1,1}, D_{2,2}) & \cdots & Sim(D_{1,1}, D_{2,n}) \\ Sim(D_{1,2}, D_{2,1}) & Sim(D_{1,2}, D_{2,2}) & \cdots & Sim(D_{1,2}, D_{2,n}) \\ \vdots & \vdots & \ddots & \vdots \\ Sim(D_{1,m}, D_{2,1}) & Sim(D_{1,m}, D_{2,2}) & \cdots & Sim(D_{1,m}, D_{2,n}) \end{bmatrix} \tag{1}$$

where $Sim(D_{1,i}, D_{2,j})$ ($i = 1, 2,\ldots, m$; $j = 1, 2,\ldots, n$) is the similarity between the complaint product $D_{1,i}$ in $\mathbf{D_1}$ and the complaint product $D_{2,j}$ in $\mathbf{D_2}$.

Step 2. *Calculate the value of each element of the similarity matrix of complaint products*. Now each element of the matrix $\mathbf{M_D}$ is the similarity between two concepts. It can be computed through using an arbitrary concept-level (word-level) measure in (Taieb et al., 2014). For instance, assume Sánchez et al.'s measure (Sánchez et al., 2012) and Lin's measure (Lin, 1998) are two candidate measures that will be chose to calculate each element of the matrix $\mathbf{M_D}$. Then:

$$Sim_{\mathbf{Sanchez}}(D_{1,i}, D_{2,j}) = 1 - \log_2\left(1 + \frac{\left|\phi(D_{1,i}) \setminus \phi(D_{2,j})\right| + \left|\phi(D_{2,j}) \setminus \phi(D_{1,i})\right|}{\left|\phi(D_{1,i}) \setminus \phi(D_{2,j})\right| + \left|\phi(D_{2,j}) \setminus \phi(D_{1,i})\right| + \left|\phi(D_{1,i}) \cap \phi(D_{2,j})\right|}\right) \tag{2}$$

$$Sim_{\mathbf{Lin}}\left(D_{1,i}, D_{2,j}\right) = \frac{2 f_{\mathbf{IC}}(\mathrm{MICA}(D_{1,i}, D_{2,j}))}{f_{\mathbf{IC}}(D_{1,i}) + f_{\mathbf{IC}}(D_{2,j})} \tag{3}$$

where $\phi(D_{1,i}) = \{C \mid C \equiv D_{1,i} \text{ or } C \sqsupseteq D_{1,i}\}$, $\phi(D_{2,j}) = \{C \mid C \equiv D_{2,j} \text{ or } C \sqsupseteq D_{2,j}\}$, $\phi(D_{1,i}) \backslash \phi(D_{2,j})$ is the set of concepts in $\phi(D_{1,i})$ but not in $\phi(D_{2,j})$ (the reverse for $\phi(D_{2,j}) \backslash \phi(D_{1,i})$), $\mathrm{MICA}(D_{1,i}, D_{2,j})$ is the most informative common ancestor of $D_{1,i}$ and $D_{2,j}$, and $f_{\mathbf{IC}}(C)$ ($C \in \{\mathrm{MICA}(D_{1,i}, D_{2,j}), D_{1,i}, D_{2,j}\}$) is a function for computing the information content of the concept C and can be computed by the following expression:

$$f_{\mathbf{IC}}(C) = -\log\left(\frac{\left|\mathrm{leaves}(C)\right|/\left|\mathrm{subsumers}(C)\right| + 1}{\mathrm{num\_of\_all\_leaves} + 1}\right) \tag{4}$$

where |leaves(C)| is the number of all terminal concepts (leaves) subsumed by C, |subsumers(C)| is the number of C itself and all its subsumers (ancestors), and num_of_all_leaves is the number of all terminal concepts (leaves) in the online complaint handling ontology.

Step 3. *Compute the similarity between the sets of complaint products*. This computation firstly selects the element $Sim_{\mathbf{Sanchez}}(D_{1,i}, D_{2,j})$ ($Sim_{\mathbf{Lin}}(D_{1,i}, D_{2,j})$) that has the maximum value from the matrix $\mathbf{M_D}$ and adds this element to a set $\mathbf{S_{Sanchez}}$ ($\mathbf{S_{Lin}}$). Then it removes all of the i-row and j-column elements of $\mathbf{M_D}$ and a new matrix $\tilde{\mathbf{M}}_\mathbf{D}$ is thus obtained. The same operations are repeated on $\tilde{\mathbf{M}}_\mathbf{D}$ until $|\mathbf{S_{Sanchez}}| = \min(m, n)$ ($|\mathbf{S_{Lin}}| = \min(m, n)$). When all operations are completed, the set $\mathbf{S_{Sanchez}}$ ($\mathbf{S_{Lin}}$) will contain a number of elements. Here assume $\mathbf{S_{Sanchez}} = \{Sim_{\mathbf{Sanchez},1}(D_{1,i}, D_{2,j}), Sim_{\mathbf{Sanchez},2}(D_{1,i}, D_{2,j}),\ldots, Sim_{\mathbf{Sanchez},N}(D_{1,i}, D_{2,j})\}$ ($\mathbf{S_{Lin}} = \{Sim_{\mathbf{Lin},1}(D_{1,i}, D_{2,j}), Sim_{\mathbf{Lin},2}(D_{1,i}, D_{2,j}),\ldots, Sim_{\mathbf{Lin},N}(D_{1,i}, D_{2,j})\}$). Then the similarity between the sets $\mathbf{D_1}$ and $\mathbf{D_2}$ can be calculated using the following expression:

$$Sim_{\mathbf{Sanchez}}(\mathbf{D_1}, \mathbf{D_2}) = \frac{m+n}{2mn} \sum_{k=1}^{N} Sim_{\mathbf{Sanchez},k}(D_{1,i}, D_{2,j}) \quad \left(Sim_{\mathbf{Lin}}(\mathbf{D_1}, \mathbf{D_2}) = \frac{m+n}{2mn} \sum_{k=1}^{N} Sim_{\mathbf{Lin},k}(D_{1,i}, D_{2,j})\right) \tag{5}$$

Similarly, the similarity between the sets $\mathbf{P_1}$ and $\mathbf{P_2}$ can be computed by establishing a similarity matrix of the complaint problems $\mathbf{M_P}$, calculating the value of each element of $\mathbf{M_P}$ using an arbitrary concept-level

(word-level) measure (e.g. $Sim_{\mathbf{Sanchez}}(P_{1,p}, P_{2,q})$ or $Sim_{\mathbf{Lin}}(P_{1,p}, P_{2,q})$ ($p = 1, 2,\ldots, u$; $q = 1, 2,\ldots, v$)), and constructing an expression to assess the similarity between $\mathbf{P}_1$ and $\mathbf{P}_2$ (e.g. $Sim_{\mathbf{Sanchez}}(\mathbf{P_1}, \mathbf{P_2})$ or $Sim_{\mathbf{Lin}}(\mathbf{P_1}, \mathbf{P_2})$).

3.3. Simlarity between contents

Online complaint contents mostly exist in the form of short texts. The similarity between the complaint contents in two online complaint cases is actually the similarity between two short texts. Generally, one needs to remove noise words like stop words, numbers, symbols, and special characters from two short texts before assessing their similarities. Through such preprocessing, the two short texts are transformed into two word sets and the similarity between two short texts is transformed into the similarity between two word sets. Formally, let $\mathbf{C}_1 = (N_1, D_1, C_1, \mathbf{D}_1, \mathbf{P}_1, T_1, \mathbf{S}_1)$, $\mathbf{C}_2 = (N_2, D_2, C_2, \mathbf{D}_2, \mathbf{P}_2, T_2, \mathbf{S}_2)$, $\mathbf{W}_1 = \{W_{1,1}, W_{1,2},\ldots, W_{1,r}\}$ be a word set that is obtained from pretreating the complaint content T_1, and $\mathbf{W}_2 = \{W_{2,1}, W_{2,2},\ldots, W_{2,s}\}$ be a word set that is obtained from pretreating the complaint content T_2. Then the similarity between T_1 and T_2 (i.e. $Sim(T_1, T_2)$) is transformed into the similarity between the word sets $\mathbf{W}_1$ and $\mathbf{W}_2$ (i.e. $Sim(\mathbf{W}_1, \mathbf{W}_2)$).

For the computation of the similarity between the word sets $\mathbf{W}_1$ and $\mathbf{W}_2$, there are two categories of methods that can be used. One is knowledge-based method and the other is corpus-based method. The similarity between $\mathbf{W}_1$ and $\mathbf{W}_2$ can be calculated by the following four steps if the knowledge-based method is chose:

Step 1. *Remove the intersection of two word sets*. Using a comparison procedure, the intersection of $\mathbf{W}_1$ and $\mathbf{W}_2$ is removed and two new word sets $\hat{\mathbf{W}}_1$ and $\hat{\mathbf{W}}_2$ are obtained. Assume the cardinality of this intersection is t, then $\hat{\mathbf{W}}_1 = \{\hat{W}_{1,1}, \hat{W}_{1,2},\ldots,\hat{W}_{1,r-t}\}$ and $\hat{\mathbf{W}}_2 = \{\hat{W}_{2,1}, \hat{W}_{2,2},\ldots,\hat{W}_{2,s-t}\}$.

Step 2. *Establish a similarity matrix of the two word sets*. A similarity matrix between $\hat{\mathbf{W}}_1$ and $\hat{\mathbf{W}}_2$ is established as follow:

$$\mathbf{M}_{\hat{\mathbf{W}}} = \begin{bmatrix} Sim(\hat{W}_{1,1},\hat{W}_{2,1}) & Sim(\hat{W}_{1,1},\hat{W}_{2,2}) & \cdots & Sim(\hat{W}_{1,1},\hat{W}_{2,s-t}) \\ Sim(\hat{W}_{1,2},\hat{W}_{2,1}) & Sim(\hat{W}_{1,2},\hat{W}_{2,2}) & \cdots & Sim(\hat{W}_{1,2},\hat{W}_{2,s-t}) \\ \vdots & \vdots & \ddots & \vdots \\ Sim(\hat{W}_{1,r-t},\hat{W}_{2,1}) & Sim(\hat{W}_{1,r-t},\hat{W}_{2,2}) & \cdots & Sim(\hat{W}_{1,r-t},\hat{W}_{2,s-t}) \end{bmatrix} \tag{6}$$

where $Sim(\hat{W}_{1,i},\hat{W}_{2,j})$ ($i = 1, 2,\ldots, r–t$; $j = 1, 2,\ldots, s–t$) is the similarity between the word $\hat{W}_{1,i}$ in $\hat{\mathbf{W}}_1$ and the word $\hat{W}_{2,j}$ in $\hat{\mathbf{W}}_2$.

Step 3. *Calculate the value of each element of the similarity matrix*. Using an arbitrary measure based on a lexical database such as WordNet (Fellbaum, 1998) or HowNet (Dong and Dong, 2006), the value of each element of the similarity matrix $\mathbf{M}_{\hat{\mathbf{W}}}$ can be computed. For example, one can choose Wu and Palmer's measure (Wu and Palmer, 1994) that is based on WordNet to calculate the value of each element if the complaint contents are in the English language. Here the calculated result is denoted as $Sim_{\mathbf{Wu}}(\hat{W}_{1,i},\hat{W}_{2,j})$, where $i = 1, 2,\ldots, r–t$ and $j = 1, 2,\ldots, s–t$.

Step 4. *Compute the similarity between the two word sets*. This computation firstly selects the element $Sim_{\mathbf{Wu}}(\hat{W}_{1,i},\hat{W}_{2,j})$ that has the maximum value from the matrix $\mathbf{M}_{\hat{\mathbf{W}}}$ and adds this element to a set $\mathbf{S}_{\mathbf{Wu}}$. Then it removes all of the i-row and j-column elements of $\mathbf{M}_{\hat{\mathbf{W}}}$ and so a new matrix $\tilde{\mathbf{M}}_{\hat{\mathbf{W}}}$ is obtained. The same operations are repeated on $\tilde{\mathbf{M}}_{\hat{\mathbf{W}}}$ until $|\mathbf{S}_{\mathbf{Wu}}| = \min(r, s) - t$. When all operations are completed, the set $\mathbf{S}_{\mathbf{Wu}}$ will contain a number of elements. Here assume $\mathbf{S}_{\mathbf{Wu}} = \{Sim_{\mathbf{Wu},1}(\hat{W}_{1,i},\hat{W}_{2,j}), Sim_{\mathbf{Wu},2}(\hat{W}_{1,i},\hat{W}_{2,j}),\ldots, Sim_{\mathbf{Wu},N}(\hat{W}_{1,i},\hat{W}_{2,j})\}$.

Then the similarity of the complaint contents T_1 and T_2 can be calculated using Islam and Inkpen's measure (Islam and Inkpen, 2008):

$$Sim_{\mathbf{Islam}}(T_1, T_2) = Sim_{\mathbf{Islam}}(\hat{\mathbf{W}}_1, \hat{\mathbf{W}}_2) = \frac{r+s}{2rs}\left(t + \sum_{k=1}^{N} Sim_{\mathbf{Wu},k}(\hat{W}_{1,i}, \hat{W}_{2,j})\right) \tag{7}$$

Correspondingly, the similarity between $\mathbf{W}_1$ and $\mathbf{W}_2$ can be computed through the following four steps if the corpus-based method is selected:

Step 1. *Establish a similarity matrix of the two word sets*. A similarity matrix between $\mathbf{W}_1$ and $\mathbf{W}_2$ is established as follow:

$$\mathbf{M_W} = \begin{bmatrix} Sim(W_{1,1}, W_{2,1}) & Sim(W_{1,1}, W_{2,2}) & \cdots & Sim(W_{1,1}, W_{2,s}) \\ Sim(W_{1,2}, W_{2,1}) & Sim(W_{1,2}, W_{2,2}) & \cdots & Sim(W_{1,2}, W_{2,s}) \\ \vdots & \vdots & \ddots & \vdots \\ Sim(W_{1,r}, W_{2,1}) & Sim(W_{1,r}, W_{2,2}) & \cdots & Sim(W_{1,r}, W_{2,s}) \end{bmatrix} \tag{8}$$

Step 2. *Calculate the value of each element of the similarity matrix*. Using an arbitrary measure based on corpus, the value of each element of the similarity matrix $\mathbf{M_W}$ can be computed. As an example, the value of each element is calculated using the following expression if Turney's measure (Turney, 2001) is chose:

$$Sim(W_{1,i}, W_{2,j}) = \frac{\text{num\_of\_lines\_hit}(W_{1,i} \& W_{2,j})}{\text{num\_of\_lines\_hit}(W_{2,j})} \tag{9}$$

where num_of_lines_hit(W) is the number of the lines in the constructed online complaint corpus which contain the word W.

Step 3. *Compute the similarity between the two word sets*. This computation firstly finds the maximum value of each row of the similarity matrix $\mathbf{M_W}$, then finds the maximum value of each column of the similarity matrix $\mathbf{M_W}$, and finally uses the following measure presented by Mihalcea et al. (2006) to calculate the similarity between the word sets $\mathbf{W}_1$ and $\mathbf{W}_2$:

$$Sim_{\mathbf{Mihalcea}}(T_1, T_2) = Sim_{\mathbf{Mihalcea}}(\mathbf{W}_1, \mathbf{W}_2) = \frac{1}{2}\left[\frac{\sum_{i=1}^{r}\left(\max_i \text{idf}(W_i)\right)}{\sum_{i=1}^{r}\text{idf}(W_i)} + \frac{\sum_{j=1}^{s}\left(\max_j \text{idf}(W_j)\right)}{\sum_{j=1}^{s}\text{idf}(W_j)}\right] \tag{10}$$

where $\max_i$ is the maximum value of the i-th row, $\max_j$ is the maximum value of the j-th column, $W_i \in \mathbf{W}_2$ is the counterpart word of $W_{1,i}$ when $Sim(W_{1,i}, W_i)$ is the maximum value of the i-th row, $W_j \in \mathbf{W}_1$ is the counterpart word of $W_{2,j}$ when $Sim(W_j, W_{2,j})$ is the maximum value of the j-th column, and idf(W) is the inverse document frequency of the word W and is computed using the online complaint corpus.

3.4. Overall similarity between cases

It is common practice for a case to be represented by a set of features. With such representation, a local similarity measure is often designed for each feature and a global similarity measure is defined as a weighted sum of all local similarity measures (De Mantaras et al., 2005). In the proposed similarity-based approach, an online complaint case is represented using seven features of case number, last revised date, complainant, complaint products, complaint problems, complaint content, and solutions, where complaint products, complaint problems, and complaint content are three distinctive features. The similarity between each of these

three features of two online complaint cases directly determines the similarity of these two online complaint cases. So far, three local similarity measures for respectively assessing the similarities between the complaint products, complaint problems, and complaint contents of two online complaint cases have been designed. A global similarity measure for assessing the overall similarity between these two cases can be naturally defined as a weighted sum of all these three local similarity measures.

Formally, let $\mathbf{C}_1 = (N_1, D_1, C_1, \mathbf{D}_1, \mathbf{P}_1, T_1, \mathbf{S}_1)$, $\mathbf{C}_2 = (N_2, D_2, C_2, \mathbf{D}_2, \mathbf{P}_2, T_2, \mathbf{S}_2)$, $Sim(\mathbf{D}_1, \mathbf{D}_2)$ be the similarity between the complaint products of the cases $\mathbf{C}_1$ and $\mathbf{C}_2$, $Sim(\mathbf{P}_1, \mathbf{P}_2)$ be the similarity between the complaint problems of $\mathbf{C}_1$ and $\mathbf{C}_2$, and $Sim(T_1, T_2)$ be the similarity between the complaint contents of $\mathbf{C}_1$ and $\mathbf{C}_2$. Then the overall similarity between $\mathbf{C}_1$ and $\mathbf{C}_2$ is defined as:

$$Sim(\mathbf{C}_1, \mathbf{C}_2) = w_1 Sim(\mathbf{D}_1, \mathbf{D}_2) + w_2 Sim(\mathbf{P}_1, \mathbf{P}_2) + w_3 Sim(T_1, T_2) \tag{11}$$

where w_1, w_2, and w_3 are respectively the weights of $Sim(\mathbf{D}_1, \mathbf{D}_2)$, $Sim(\mathbf{P}_1, \mathbf{P}_2)$, and $Sim(T_1, T_2)$ such that $0 \leq w_1, w_2, w_3 \leq 1$ and $w_1 + w_2 + w_3 = 1$. As can be seen from this expression, the value of $Sim(\mathbf{C}_1, \mathbf{C}_2)$ is determined by the values of w_1, w_2, and w_3 when the values of $Sim(\mathbf{D}_1, \mathbf{D}_2)$, $Sim(\mathbf{P}_1, \mathbf{P}_2)$, and $Sim(T_1, T_2)$ are certain. That is to say, the assignment of w_1, w_2, and w_3 directly affects the similarity assessment accuracy of $Sim(\mathbf{C}_1, \mathbf{C}_2)$ and will have an influence on the accuracy of case retrieval. Thus, how to assign w_1, w_2, and w_3 that can ensure the similarity assessment accuracy of $Sim(\mathbf{C}_1, \mathbf{C}_2)$ is the first question needed to be addressed.

Generally, the similarity assessment accuracy of a measure can be quantified by the Pearson correlation coefficient between the similarities of a certain number of randomly selected sample pairs computed by the measure and the similarities of these sample pairs judged by a certain number of domain experts (the similarity of each sample pair is the mean value of the similarities of this sample pair judged by a certain number of domain experts). The greater this Pearson correlation coefficient, the higher the similarity assessment accuracy of this measure is (Taieb et al., 2014; Pilehvar and Navigli, 2015). Hence, the values of w_1, w_2, and w_3 can be calculated out by maximizing the Pearson correlation coefficient between the similarities of a certain number of randomly selected sample case pairs computed by $Sim(\mathbf{C}_1, \mathbf{C}_2)$ and the similarities of these sample case pairs judged by a certain number of domain experts.

Formally, let $Sim_\mathbf{J}(\mathbf{C}_1, \mathbf{C}_2)$ be the similarity of $\mathbf{C}_1$ and $\mathbf{C}_2$ judged by a certain number of domain experts, vector $A = [Sim(\mathbf{D}_1, \mathbf{D}_2), Sim(\mathbf{P}_1, \mathbf{P}_2), Sim(T_1, T_2)]^\mathrm{T}$, vector $B = [Sim_\mathbf{J}(\mathbf{C}_1, \mathbf{C}_2)]^\mathrm{T}$, and vector $\hat{w} = [\hat{w}_1, \hat{w}_2, \hat{w}_3]^\mathrm{T}$. Then the question of assigning w_1, w_2, and w_3 can be transformed to a question of seeking the vector $\hat{w}$ that can maximize the Pearson correlation coefficient between $\hat{w}^\mathrm{T} A$ and B:

$$\rho = \mathrm{corr}(\hat{w}^\mathrm{T} A, B) = \frac{\hat{w}^\mathrm{T} \mathrm{cov}(A, B)}{\sqrt{\hat{w}^\mathrm{T} \mathrm{cov}(A, A) \hat{w}} \sqrt{\mathrm{cov}(B, B)}} \tag{12}$$

where cov is the abbreviation of covariance and cov(A, A), cov(A, B), and cov(B, B) are respectively the following 3×3, 3×1, and 1×1 matrices:

$$\mathrm{cov}(A, A) = \begin{bmatrix} \mathrm{cov}(Sim(\mathbf{D}_1, \mathbf{D}_2), Sim(\mathbf{D}_1, \mathbf{D}_2)) & \mathrm{cov}(Sim(\mathbf{D}_1, \mathbf{D}_2), Sim(\mathbf{P}_1, \mathbf{P}_2)) & \mathrm{cov}(Sim(\mathbf{D}_1, \mathbf{D}_2), Sim(T_1, T_2)) \\ \mathrm{cov}(Sim(\mathbf{P}_1, \mathbf{P}_2), Sim(\mathbf{D}_1, \mathbf{D}_2)) & \mathrm{cov}(Sim(\mathbf{P}_1, \mathbf{P}_2), Sim(\mathbf{P}_1, \mathbf{P}_2)) & \mathrm{cov}(Sim(\mathbf{P}_1, \mathbf{P}_2), Sim(T_1, T_2)) \\ \mathrm{cov}(Sim(T_1, T_2), Sim(\mathbf{D}_1, \mathbf{D}_2)) & \mathrm{cov}(Sim(T_1, T_2), Sim(\mathbf{P}_1, \mathbf{P}_2)) & \mathrm{cov}(Sim(T_1, T_2), Sim(T_1, T_2)) \end{bmatrix} \tag{13}$$

$$\text{cov}(A, B) = \begin{bmatrix} \text{cov}(Sim(\mathbf{D}_1, \mathbf{D}_2), Sim_{\mathbf{J}}(\mathbf{C}_1, \mathbf{C}_2)) \\ \text{cov}(Sim(\mathbf{P}_1, \mathbf{P}_2), Sim_{\mathbf{J}}(\mathbf{C}_1, \mathbf{C}_2)) \\ \text{cov}(Sim(T_1, T_2), Sim_{\mathbf{J}}(\mathbf{C}_1, \mathbf{C}_2)) \end{bmatrix} \tag{14}$$

$$\text{cov}(B, B) = \left[\text{cov}(Sim_{\mathbf{J}}(\mathbf{C}_1, \mathbf{C}_2)), Sim_{\mathbf{J}}(\mathbf{C}_1, \mathbf{C}_2)) \right] \tag{15}$$

To solve the vector $\hat{w}$, a canonical correlation analysis method (Lu et al., 2016) is used. The general solving process is as follows. Firstly, let vector $u = \left(\text{cov}(A, A) \right)^{1/2} \hat{w}$ and vector $v = \left(\text{cov}(B, B) \right)^{1/2}$. Then Expression (12) can be transformed to:

$$\rho = \left[u^{\mathrm{T}} \left(\text{cov}(A, A) \right)^{-1/2} \text{cov}(A, B) \left(\text{cov}(B, B) \right)^{-1/2} v \right] \Big/ \left(\sqrt{u^{\mathrm{T}} u} \sqrt{v^{\mathrm{T}} v} \right) = \varepsilon \Big/ \left(\sqrt{u^{\mathrm{T}} u} \sqrt{v^{\mathrm{T}} v} \right) \tag{16}$$

According to the Cauchy-Schwarz inequality, the following inequality is obtained:

$$\varepsilon \leq \sqrt{u^{\mathrm{T}} \left(\text{cov}(A, A) \right)^{-1/2} \text{cov}(A, B) \left(\text{cov}(B, B) \right)^{-1} \text{cov}(B, A) \left(\text{cov}(A, A) \right)^{-1/2} u} \sqrt{v^{\mathrm{T}} v} \tag{17}$$

where $\text{cov}(B, A)$ is the following 1×3 matrix:

$$\text{cov}(B, A) = \begin{bmatrix} \text{cov}(Sim_{\mathbf{J}}(\mathbf{C}_1, \mathbf{C}_2), Sim(\mathbf{D}_1, \mathbf{D}_2)) \\ \text{cov}(Sim_{\mathbf{J}}(\mathbf{C}_1, \mathbf{C}_2), Sim(\mathbf{P}_1, \mathbf{P}_2)) \\ \text{cov}(Sim_{\mathbf{J}}(\mathbf{C}_1, \mathbf{C}_2), Sim(T_1, T_2)) \end{bmatrix}^{\mathrm{T}} \tag{18}$$

Thus the following inequality holds:

$$\rho \leq \sqrt{u^{\mathrm{T}} \left(\text{cov}(A, A) \right)^{-1/2} \text{cov}(A, B) \left(\text{cov}(B, B) \right)^{-1} \text{cov}(B, A) \left(\text{cov}(A, A) \right)^{-1/2} u} \Big/ \sqrt{u^{\mathrm{T}} u} \tag{19}$$

As can be seen from Expression (19), the maximum of ρ is attained if u is the eigenvector with the maximum eigenvalue for the matrix $\left(\text{cov}(A, A) \right)^{-1/2} \text{cov}(A, B) \left(\text{cov}(B, B) \right)^{-1} \text{cov}(B, A) \left(\text{cov}(A, A) \right)^{-1/2}$. Thus the solution is: $\hat{w}$ is an eigenvector with the maximum eigenvalue for $\left(\text{cov}(A, A) \right)^{-1} \text{cov}(A, B) \left(\text{cov}(B, B) \right)^{-1} \text{cov}(B, A)$. If N sample case pairs that are randomly selected from the online complaint case base and their judged similarities are given, the values of all elements of $\hat{w}$ (i.e. $\hat{w}_1, \hat{w}_2,$ and $\hat{w}_3$) will be calculated out. However, these values are not the final weights of $Sim(\mathbf{D}_1, \mathbf{D}_2)$, $Sim(\mathbf{P}_1, \mathbf{P}_2)$, and $Sim(T_1, T_2)$ since their sum is usually not equal to 1. By a normalization, the final weights (i.e. $w_1, w_2,$ and w_3) are obtained as: $w_i = \hat{w}_i / (\hat{w}_1 + \hat{w}_2 + \hat{w}_3)$ (i = 1, 2, 3).

As can be seen from Expression (11) and the above weight solving process, the value of $Sim(\mathbf{C}_1, \mathbf{C}_2)$ and the values of $w_1, w_2,$ and w_3 are all determined by the values of $Sim(\mathbf{D}_1, \mathbf{D}_2)$, $Sim(\mathbf{P}_1, \mathbf{P}_2)$, and $Sim(T_1, T_2)$. That is to say, how to respectively choose three measures to calculate out the values of $Sim(\mathbf{D}_1, \mathbf{D}_2)$, $Sim(\mathbf{P}_1, \mathbf{P}_2)$, and $Sim(T_1, T_2)$ also directly affects the similarity assessment accuracy of $Sim(\mathbf{C}_1, \mathbf{C}_2)$ and will also have an influence on the accuracy of case retrieval. So how to respectively choose three measures that can ensure the similarity assessment accuracy of $Sim(\mathbf{C}_1, \mathbf{C}_2)$ is the second question required to be addressed.

Naturally, the three measures can also be chosen through maximizing the Pearson correlation coefficient between the similarities of a certain number of randomly selected sample case pairs computed by $Sim(\mathbf{C}_1, \mathbf{C}_2)$ and the similarities of these sample case pairs judged by a certain number of domain experts. Formally, assume $Sim_1(\mathbf{D}_1, \mathbf{D}_2)$, $Sim_2(\mathbf{D}_1, \mathbf{D}_2), \ldots, Sim_K(\mathbf{D}_1, \mathbf{D}_2)$ are K candidate measures that can be chose to calculate out the value of $Sim(\mathbf{D}_1, \mathbf{D}_2)$, $Sim_1(\mathbf{P}_1, \mathbf{P}_2)$, $Sim_2(\mathbf{P}_1, \mathbf{P}_2), \ldots, Sim_L(\mathbf{P}_1, \mathbf{P}_2)$ are L candidate measures that can be

chose to calculate out the value of $Sim(\mathbf{P}_1, \mathbf{P}_2)$, and $Sim_1(T_1, T_2)$, $Sim_2(T_1, T_2)$,…, $Sim_M(T_1, T_2)$ are M candidate measures that can be selected to calculate out the value of $Sim(T_1, T_2)$. Then $K{\times}L{\times}M$ measures for calculating out the values of $Sim(\mathbf{C}_1, \mathbf{C}_2)$ ($Sim_1(\mathbf{C}_1, \mathbf{C}_2)$, $Sim_2(\mathbf{C}_1, \mathbf{C}_2)$,…, $Sim_{K{\times}L{\times}M}(\mathbf{C}_1, \mathbf{C}_2)$) will be obtained. If N sample case pairs that are randomly selected from the case base and their judged similarities are given, the values of each group of weights, the values of each of the $K{\times}L{\times}M$ similarities for each sample case pair, and the Pearson correlation coefficient between each of the $K{\times}L{\times}M$ similarities and the judged similarities will be successively calculated out. Then one of the $K{\times}L{\times}M$ measures that can obtain the greatest Pearson correlation coefficient will be chose as the final measure to assess the similarity of a target case and each of all source cases.

For example, if $Sim_{\mathbf{Sanchez}}(\mathbf{D}_1, \mathbf{D}_2)$ (Expression (5)) and $Sim_{\mathbf{Lin}}(\mathbf{D}_1, \mathbf{D}_2)$ (Expression (5)) are two candidate measures for $Sim(\mathbf{D}_1, \mathbf{D}_2)$, $Sim_{\mathbf{Sanchez}}(\mathbf{P}_1, \mathbf{P}_2)$ (similar to Expression (5)) and $Sim_{\mathbf{Lin}}(\mathbf{P}_1, \mathbf{P}_2)$ (similar to Expression (5)) are two candidate measures for $Sim(\mathbf{P}_1, \mathbf{P}_2)$, and $Sim_{\mathbf{Islam}}(T_1, T_2)$ (Expression (7)) and $Sim_{\mathbf{Mihalcea}}(T_1, T_2)$ (Expression (10)) are two candidate measures for $Sim(T_1, T_2)$, then 8 measures for calculating out the values of $Sim(\mathbf{C}_1, \mathbf{C}_2)$ will be obtained:

$$
\begin{cases}
Sim_1(\mathbf{C}_1, \mathbf{C}_2) = w_{1,1}Sim_{\mathbf{Sanchez}}(\mathbf{D}_1, \mathbf{D}_2) + w_{1,2}Sim_{\mathbf{Sanchez}}(\mathbf{P}_1, \mathbf{P}_2) + w_{1,3}Sim_{\mathbf{Islam}}(T_1, T_2) \\
Sim_2(\mathbf{C}_1, \mathbf{C}_2) = w_{2,1}Sim_{\mathbf{Sanchez}}(\mathbf{D}_1, \mathbf{D}_2) + w_{2,2}Sim_{\mathbf{Sanchez}}(\mathbf{P}_1, \mathbf{P}_2) + w_{2,3}Sim_{\mathbf{Mihalcea}}(T_1, T_2) \\
Sim_3(\mathbf{C}_1, \mathbf{C}_2) = w_{3,1}Sim_{\mathbf{Sanchez}}(\mathbf{D}_1, \mathbf{D}_2) + w_{3,2}Sim_{\mathbf{Lin}}(\mathbf{P}_1, \mathbf{P}_2) + w_{3,3}Sim_{\mathbf{Islam}}(T_1, T_2) \\
Sim_4(\mathbf{C}_1, \mathbf{C}_2) = w_{4,1}Sim_{\mathbf{Sanchez}}(\mathbf{D}_1, \mathbf{D}_2) + w_{4,2}Sim_{\mathbf{Lin}}(\mathbf{P}_1, \mathbf{P}_2) + w_{4,3}Sim_{\mathbf{Mihalcea}}(T_1, T_2) \\
Sim_5(\mathbf{C}_1, \mathbf{C}_2) = w_{5,1}Sim_{\mathbf{Lin}}(\mathbf{D}_1, \mathbf{D}_2) + w_{5,2}Sim_{\mathbf{Sanchez}}(\mathbf{P}_1, \mathbf{P}_2) + w_{5,3}Sim_{\mathbf{Islam}}(T_1, T_2) \\
Sim_6(\mathbf{C}_1, \mathbf{C}_2) = w_{6,1}Sim_{\mathbf{Lin}}(\mathbf{D}_1, \mathbf{D}_2) + w_{6,2}Sim_{\mathbf{Sanchez}}(\mathbf{P}_1, \mathbf{P}_2) + w_{6,3}Sim_{\mathbf{Mihalcea}}(T_1, T_2) \\
Sim_7(\mathbf{C}_1, \mathbf{C}_2) = w_{7,1}Sim_{\mathbf{Lin}}(\mathbf{D}_1, \mathbf{D}_2) + w_{7,2}Sim_{\mathbf{Lin}}(\mathbf{P}_1, \mathbf{P}_2) + w_{7,3}Sim_{\mathbf{Islam}}(T_1, T_2) \\
Sim_8(\mathbf{C}_1, \mathbf{C}_2) = w_{8,1}Sim_{\mathbf{Lin}}(\mathbf{D}_1, \mathbf{D}_2) + w_{8,2}Sim_{\mathbf{Lin}}(\mathbf{P}_1, \mathbf{P}_2) + w_{8,3}Sim_{\mathbf{Mihalcea}}(T_1, T_2)
\end{cases} \tag{20}
$$

Now if N sample case pairs $(\mathbf{C}_{i,1}, \mathbf{C}_{i,2})$ ($i = 1, 2,…, N$) which are randomly selected from the online complaint case base and their judged similarities $Sim_{\mathbf{J}}(\mathbf{C}_{i,1}, \mathbf{C}_{i,2})$ are given, the values of $w_{j,1}$, $w_{j,2}$, and $w_{j,3}$ ($j = 1, 2,…, 8$), $Sim_j(\mathbf{C}_{i,1}, \mathbf{C}_{i,2})$, and $corr(Sim_j(\mathbf{C}_{i,1}, \mathbf{C}_{i,2}), Sim_{\mathbf{J}}(\mathbf{C}_{i,1}, \mathbf{C}_{i,2}))$ will be successively calculated out. Then one of the 8 measures $Sim_j(\mathbf{C}_1, \mathbf{C}_2)$ that can obtain the greatest Pearson correlation coefficient will be selected as the final measure to compute the similarity between a new target case and each of all source cases.

4. Implementation and evaluation

This section firstly reports a prototype implementation of the proposed similarity-based approach. It then evaluates the approach through numerical and empirical experiments.

4.1. Implementation

The online complaint handling ontology was developed using Protégé, a free and open-source ontology editor providing an integration environment of creating, editing, and saving OWL ontologies in a visual way. The online complaint corpus was developed through using Octopus, an Internet data fetching, processing, and mining software, to fetch 5000 online complaints about the products of AT&T from Better Business Bureau, preprocessing the fetched online complaints to obtain a set of words for each online complaint, and storing this set of words to a separate line in the same text document. The computation of the similarities between online complaint cases was implemented with the use of Protégé-OWL API (application program interface), WordNet Java API, and the Java programming language.

4.2. Evaluation

4.2.1. Numerical experiments

Generally, a numerical experiment for evaluating the accuracies of different measures can be made using an identical benchmark which consists of a certain number of sense pairs, word pairs, or text pairs and their judged similarities. During the past few decades, a number of benchmarks at different linguistic levels, which include sense, word, and text levels, have been designed (Pilehvar and Navigli, 2015). Evaluating the accuracies of different measures for assessing the similarity between online complaint cases requires an identical text-level benchmark since online complaints mostly exist in the form of short texts. But existing text-level benchmarks cannot satisfy this requirement because their contents are not online complaints. A specialized text-level benchmark in the field of online complaint handling is required to be designed.

To design the benchmark, 100 sample case pairs were randomly selected from the constructed case base containing 500 source cases. Then the similarity of each sample case pair on a scale 0, 0.2, 0.4, 0.6, 0.8, and 1 was judged by 12 communication services' regular customers. The 12 judged results of each sample case pair were sorted in descending order and the first one and last one results were removed. The mean value of the rest 10 results was calculated and considered as the judged similarity of each sample case pair.

According to the approach, the determined similarity measure is one of the eight similarity measures in Expression (20). Because each of these eight similarity measures has different linear combinations, eight numerical experiments (each of the eight similarity measures corresponds to a numerical experiment) are required to be conducted to verify that the similarity measure determined by the proposed approach is more accurate than all of its linear combinations. In the eight numerical experiments, the benchmark was used to evaluate the accuracies of the measures considering only one feature, the measures considering two features, the measure considering three features, and the measure determined by the proposed approach.

Specifically, each experiment was carried out by comparing the accuracies of all possible linear combinations of the three components in each of these eight measures in Expression (20). For example, the first experiment was carried out through comparing the accuracies of all possible linear combinations of the three components $Sim_{\mathbf{Sanchez}}(\mathbf{D_1}, \mathbf{D_2})$, $Sim_{\mathbf{Sanchez}}(\mathbf{P_1}, \mathbf{P_2})$, and $Sim_{\mathbf{Islam}}(T_1, T_2)$ in the first measure $Sim_1(\mathbf{C_1}, \mathbf{C_2})$ in Expression (20). Such linear combinations are listed as follows:

$$Sim_{1,1}(\mathbf{C_1}, \mathbf{C_2}) = Sim_{\mathbf{Sanchez}}(\mathbf{D_1}, \mathbf{D_2}) \tag{21}$$

$$Sim_{1,2}(\mathbf{C_1}, \mathbf{C_2}) = Sim_{\mathbf{Sanchez}}(\mathbf{P_1}, \mathbf{P_2}) \tag{22}$$

$$Sim_{1,3}(\mathbf{C_1}, \mathbf{C_2}) = Sim_{\mathbf{Islam}}(T_1, T_2) \tag{23}$$

$$Sim_{2,1}(\mathbf{C_1}, \mathbf{C_2}) = w_{2,1,1}Sim_{\mathbf{Sanchez}}(\mathbf{D_1}, \mathbf{D_2}) + w_{2,1,2}Sim_{\mathbf{Sanchez}}(\mathbf{P_1}, \mathbf{P_2}) \tag{24}$$

$$Sim_{2,2}(\mathbf{C_1}, \mathbf{C_2}) = w_{2,2,1}Sim_{\mathbf{Sanchez}}(\mathbf{D_1}, \mathbf{D_2}) + w_{2,2,2}Sim_{\mathbf{Islam}}(T_1, T_2) \tag{25}$$

$$Sim_{2,3}(\mathbf{C_1}, \mathbf{C_2}) = w_{2,3,1}Sim_{\mathbf{Sanchez}}(\mathbf{P_1}, \mathbf{P_2}) + w_{2,3,2}Sim_{\mathbf{Islam}}(T_1, T_2) \tag{26}$$

$$Sim_{3,1}(\mathbf{C_1}, \mathbf{C_2}) = w_{3,1,1}Sim_{\mathbf{Sanchez}}(\mathbf{D_1}, \mathbf{D_2}) + w_{3,1,2}Sim_{\mathbf{Sanchez}}(\mathbf{P_1}, \mathbf{P_2}) + w_{3,1,3}Sim_{\mathbf{Islam}}(T_1, T_2) \tag{27}$$

The experiment firstly used the weight solving procedure to calculate out the values of each group of weights in $Sim_{2,1}(\mathbf{C_1}, \mathbf{C_2})$, $Sim_{2,2}(\mathbf{C_1}, \mathbf{C_2})$, $Sim_{2,3}(\mathbf{C_1}, \mathbf{C_2})$, and $Sim_{3,1}(\mathbf{C_1}, \mathbf{C_2})$. It then used each $Sim_{j,k}(\mathbf{C_1}, \mathbf{C_2})$ ($j = 1, 2, 3$; $k = 1, 2, 3$) to assess the similarities of the 100 sample case pairs in the benchmark and computed the correla-

tion coefficient between each group of the assessed similarities and the judged similarities of the 100 sample case pairs. The calculated each group of weights and correlation coefficient is listed in Table 1. Similarly, the remaining seven experiments were completed and their results are also listed in Table 1.

Table 1

The calculated each group of weights and Pearson correlation coefficient in each of the eight experiments. $Sim_{j,k}$ ($j = 1, 2, 3$; $k = 1, 2, 3$) is short for $Sim_{j,k}(\mathbf{C}_1, \mathbf{C}_2)$ ($Sim_{j,k}(\mathbf{C}_1, \mathbf{C}_2)$ in the h-th ($h = 1, 2,\ldots, 8$) experiment are the all possible linear combinations of the three components in $Sim_h(\mathbf{C}_1, \mathbf{C}_2)$ in Expression (20). For instance, $Sim_{j,k}(\mathbf{C}_1, \mathbf{C}_2)$ in the first experiment are the seven possible linear combinations of the three components $Sim_{\mathbf{Sanchez}}(\mathbf{D}_1, \mathbf{D}_2)$, $Sim_{\mathbf{Sanchez}}(\mathbf{P}_1, \mathbf{P}_2)$, and $Sim_{\mathbf{Islam}}(T_1, T_2)$ in $Sim_1(\mathbf{C}_1, \mathbf{C}_2)$ in Expression (20) (i.e. Expressions (21)–(27))). $corr_{j,k}$ is short for $corr(Sim_{j,k}(\mathbf{C}_{i,1}, \mathbf{C}_{i,2}), Sim_{\mathbf{J}}(\mathbf{C}_{i,1}, \mathbf{C}_{i,2}))$ ($i = 1, 2,\ldots, 100$).

Experiment	w&corr	$Sim_{1,1}$	$Sim_{1,2}$	$Sim_{1,3}$	$Sim_{2,1}$	$Sim_{2,2}$	$Sim_{2,3}$	$Sim_{3,1}$
Experiment 1	$w_{j,k,1}$	1.0000	1.0000	1.0000	0.5946	0.5764	0.6816	0.4178
	$w_{j,k,2}$	—	—	—	0.4054	0.4236	0.3184	0.2535
	$w_{j,k,3}$	—	—	—	—	—	—	0.3287
	$corr_{j,k}$	0.7333	0.6764	0.2078	0.7745	0.7523	0.6836	**0.7850**
Experiment 2	$w_{j,k,1}$	1.0000	1.0000	1.0000	0.5946	0.5737	0.4600	0.3162
	$w_{j,k,2}$	—	—	—	0.4054	0.4263	0.5400	0.2956
	$w_{j,k,3}$	—	—	—	—	—	—	0.3882
	$corr_{j,k}$	0.7333	0.6764	0.5514	0.7745	0.7652	0.7641	**0.8062**
Experiment 3	$w_{j,k,1}$	1.0000	1.0000	1.0000	0.7075	0.5764	0.5679	0.4709
	$w_{j,k,2}$	—	—	—	0.2925	0.4236	0.4321	0.1764
	$w_{j,k,3}$	—	—	—	—	—	—	0.3527
	$corr_{j,k}$	0.7333	0.5791	0.2078	0.7648	0.7523	0.5928	**0.7776**
Experiment 4	$w_{j,k,1}$	1.0000	1.0000	1.0000	0.7075	0.5737	0.3623	0.3732
	$w_{j,k,2}$	—	—	—	0.2925	0.4263	0.6377	0.2204
	$w_{j,k,3}$	—	—	—	—	—	—	0.4064
	$corr_{j,k}$	0.7333	0.5791	0.5514	0.7648	0.7652	0.7265	**0.8003**
Experiment 5	$w_{j,k,1}$	1.0000	1.0000	1.0000	0.3806	0.5205	0.6816	0.2748
	$w_{j,k,2}$	—	—	—	0.6194	0.4795	0.3184	0.4100
	$w_{j,k,3}$	—	—	—	—	—	—	0.3152
	$corr_{j,k}$	0.5980	0.6764	0.2078	0.7163	0.6251	0.6836	**0.7257**
Experiment 6	$w_{j,k,1}$	1.0000	1.0000	1.0000	0.3806	0.4252	0.4600	0.1276
	$w_{j,k,2}$	—	—	—	0.6194	0.5748	0.5400	0.3986
	$w_{j,k,3}$	—	—	—	—	—	—	0.4738
	$corr_{j,k}$	0.5980	0.6764	0.5514	0.7163	0.6707	0.7641	**0.7720**
Experiment 7	$w_{j,k,1}$	1.0000	1.0000	1.0000	0.5482	0.5205	0.5679	0.3668
	$w_{j,k,2}$	—	—	—	0.4518	0.4795	0.4321	0.2839
	$w_{j,k,3}$	—	—	—	—	—	—	0.3493
	$corr_{j,k}$	0.5980	0.5791	0.2078	0.6932	0.6251	0.5928	**0.7063**
Experiment 8	$w_{j,k,1}$	1.0000	1.0000	1.0000	0.5482	0.4252	0.3623	0.2187
	$w_{j,k,2}$	—	—	—	0.4518	0.5748	0.6377	0.2921
	$w_{j,k,3}$	—	—	—	—	—	—	0.4892
	$corr_{j,k}$	0.5980	0.5791	0.5514	0.6932	0.6707	0.7265	**0.7566**

4.2.2. Numerical experiment result analysis

As can be seen from Table 1, corr($Sim_2(\mathbf{C}_{i,1}, \mathbf{C}_{i,2})$, $Sim_\mathbf{J}(\mathbf{C}_{i,1}, \mathbf{C}_{i,2})$) (i.e. 0.8062) $\geq$ corr($Sim_{3,1}(\mathbf{C}_{i,1}, \mathbf{C}_{i,2})$, $Sim_\mathbf{J}(\mathbf{C}_{i,1}, \mathbf{C}_{i,2})$) > corr($Sim_{j,k}(\mathbf{C}_{i,1}, \mathbf{C}_{i,2})$, $Sim_\mathbf{J}(\mathbf{C}_{i,1}, \mathbf{C}_{i,2})$) for all $j = 1, 2$ and $k = 1, 2, 3$ (where $i = 1, 2, \ldots, 100$) holds in every experiment, which experimentally verifies the effectiveness of the proposed approach. Moreover, careful readers may find there are slight differences among the maximum correlation coefficients in the eight experiments. This is because they are all the optimal values obtained by using the same optimization method to optimize the accuracies of the eight global measures that differentiate with each other in only one local measure.

As can also be seen from Table 1, the measure $Sim_\mathbf{Islam}(T_1, T_2)$ has a very low human correlation (0.2078). This is probably because this measure computed the similarities of the 100 sample case pairs on the basis of WordNet which only contains conventional words and each online complaint may include a certain number of network words that cannot be found in WordNet. For this reason, it is recommended that the assessment of the similarities between online complaint cases leverages some specialized knowledge containers instead of some general lexical databases like WordNet and HowNet that only contain conventional words. Correspondingly, the human correlation of the measure $Sim_\mathbf{Mihalcea}(T_1, T_2)$ that works based on the constructed online complaint corpus can reach 0.5514. This demonstrates that the constructed corpus consisting of online complaint words is helpful. It is believed that the more perfect the corpus is, the better the performance of this measure is. Different from these two measures, there is no big difference among the human correlations of the rest four candidate measures $Sim_\mathbf{Sanchez}(\mathbf{D}_1, \mathbf{D}_2)$, $Sim_\mathbf{Lin}(\mathbf{D}_1, \mathbf{D}_2)$, $Sim_\mathbf{Sanchez}(\mathbf{P}_1, \mathbf{P}_2)$, and $Sim_\mathbf{Lin}(\mathbf{P}_1, \mathbf{P}_2)$ (whose human correlations are 0.7333, 0.5980, 0.6764, and 0.5791, respectively). This reveals that the constructed semantic networks of complaint products and problems in the online complaint handling ontology can play a good role since all these candidate measures work on the basis of such semantic networks.

Some readers may ask that some existing word-level measures (e.g. Pilehvar and Navigli's measure (Pilehvar and Navigli, 2015)) can reach a human correlation of more than 0.9 but the highest correlation among the correlations of the four word-level measures here (i.e. $Sim_\mathbf{Sanchez}(\mathbf{D}_1, \mathbf{D}_2)$, $Sim_\mathbf{Lin}(\mathbf{D}_1, \mathbf{D}_2)$, $Sim_\mathbf{Sanchez}(\mathbf{P}_1, \mathbf{P}_2)$, and $Sim_\mathbf{Lin}(\mathbf{P}_1, \mathbf{P}_2)$) is only 0.7333 and the correlation of the measure found out by the proposed approach is only 0.8062. The reasons for this phenomenon are as follows: (1) The sample word pairs in those word-level measures obtaining a human correlation of more than 0.9 are conventional word pairs that are contained in some powerful knowledge containers like WordNet and HowNet, but the sample word pairs in the four word-level measures here are some specialized word pairs about the complaint products and problems of AT&T that do not have a powerful knowledge container to support. (2) A highly accurate assessment of the similarity between texts is a challenge and difficult task in natural language processing (Pilehvar and Navigli, 2015) and unfortunately online complaints mostly exist in the form of short texts in which the string sequences and character compositions are always disorganized. Even so, it is still believed that the human correlation of the designed global measure for assessing the similarity between online complaint cases can be further improved by perfecting the constructed online complaint handling ontology and online complaint corpus and trying to use more existing measures as candidate measures to compute the local similarities.

In summary, the results and findings of the numerical experiments are as follows: (1) The measure found out by the proposed approach was verified to be more accurate than all of its linear combinations by provid-

ing the highest human correlation among the human correlations of these linear combinations. (2) The specialized knowledge containers (i.e. the constructed ontology and corpus) can play a good role in the assessment of the similarity of two online complaint cases. (3) A global measure considering the features at all linguistic levels comprehensively and selecting different local measures for different case bases can obtain more competitive accuracy than those global measures considering only one feature and using only one local measure. (4) The weight solving and measure selection procedures in the proposed approach were helpful to improve the accuracy of the global measure. (5) There is still room for improvement for the accuracy of the designed global measure for assessing the similarity between online complaint cases.

4.2.3. Empirical experiment

A retrieval of similar source cases from the constructed case base for a new target case is taken as an empirical example to verify the effectiveness of the proposed approach. This retrieval is supported by three knowledge containers: the constructed online complaint handling ontology, WordNet, and the established online complaint corpus. It takes as input 100 sample case pairs and their judged similarities (i.e. the designed benchmark), a new online complaint about the product of AT&T (i.e. a new target case "*I've signed up for **** internet service a year ago and it seems every couple of months I need to contact them regarding billing on my account. I first signed up for unlimited internet for $19.95. After a few months I started being overcharged to $57/month. After a lot of back/forth communication with AT&T, I agreed with $57 charge for the same service since they told me they didn't have that package anymore. On April/2016 I was charged $107.00 for the service which was 19.95 a year ago. I already spoke to Customer Service and they informed me the charge was because I passed my internet usage limit. I shouldn't have any limit since my service is unlimited. I need BBB help with this matter. Previous account 05/2015 Order Number: Y XXXX XXXXX **** High Speed Internet Telephone/Account Number: XXX-XXX-XXXX I don't have info on original account from 03/2015 Product_Or_Service: **** home internet service.*"), a similarity threshold 0.8, and a positive integer 8, and it will return as output 8 source cases sorted in descending order based on their similarities with the new target case. The details of such a process from input to output, which mainly contain four steps, are explained below.

The first step is to find out a measure with high accuracy for assessing the overall similarity between cases. It firstly assessed the similarities between the complaint products of the 100 sample case pairs using $Sim_{Sanchez}(\mathbf{D_1}, \mathbf{D_2})$ and $Sim_{Lin}(\mathbf{D_1}, \mathbf{D_2})$, assessed the similarities between the complaint problems of the 100 sample case pairs using $Sim_{Sanchez}(\mathbf{P_1}, \mathbf{P_2})$ and $Sim_{Lin}(\mathbf{P_1}, \mathbf{P_2})$, and assessed the similarities between the complaint contents of the 100 sample case pairs using $Sim_{Islam}(T_1, T_2)$ and $Sim_{Mihalcea}(T_1, T_2)$. Then the values of the 8 groups of weights in Expression (20) were, as listed in the last column of Table 1, respectively calculated out by the weight solving procedure in the previous section (from Expression (12) to Expression (19)). With these assessed similarities and calculated weights, the overall similarities of the 100 sample case pairs were computed out respectively using the 8 measures in Expression (20). Finally, the correlation coefficient between each group of these overall similarities and the judged similarities of the 100 sample case pairs was calculated out and is also listed in the last column of Table 1. As can be seen from this column, the correlation coefficient between $Sim_2(\mathbf{C}_{i,1}, \mathbf{C}_{i,2})$ and $Sim_{\mathbf{J}}(\mathbf{C}_{i,1}, \mathbf{C}_{i,2})$ ($i = 1, 2, \ldots, 100$) is the greatest, thus the measure with high accuracy for assessing the overall similarity between cases is determined as:

$$Sim(\mathbf{C_1}, \mathbf{C_2}) = 0.3162Sim_{\text{Sanchez}}(\mathbf{D_1}, \mathbf{D_2}) + 0.2956Sim_{\text{Sanchez}}(\mathbf{P_1}, \mathbf{P_2}) + 0.3882Sim_{\text{Mihalcea}}(T_1, T_2) \qquad (28)$$

The second step is to extract the complaint products and problems from the complaint content of the new target case. Using Gupta's extraction method (Gupta, 2013), the complaint product and problem of the new target case were automatically extracted from its complaint content. As shown in Figure 2, the extracted complaint product and problem are "*Internet*" and "*Overcharge*", respectively.
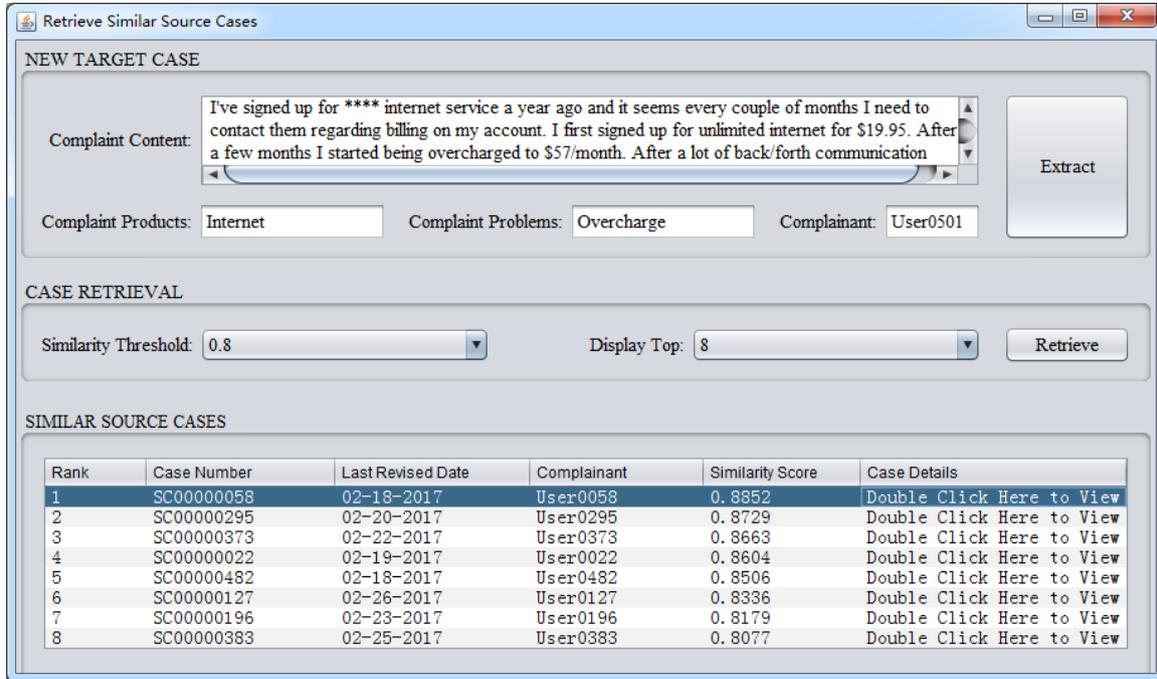
Figure 2. The top 8 source cases in all of the sorted source cases.

The third step is to retrieve similar source cases for the new target case. Using the determined measure in the first step, the similarity between the new target case and each of the 500 source cases in the case base was computed. Then all source cases whose similarities with the new target case are greater than or equal to the input similarity threshold 0.8 were screened out. After that, the screened out source cases were sorted in descending order based on their similarities with the new target case and the top 8 (8 is the input positive integer) source cases in all of the sorted source cases were output and also shown in Figure 2.

The last step is to reuse the retrieved similar source cases to handle the online complaint in the new target case. The solutions in the output 8 source cases could be used as reference solutions of the complaint problem in the new target case. For example, the solutions of the source case SC00000058 in Figure 2 (whose details are shown in Figure 3) were taken as reference solutions to handle the input new online complaint.

As can be seen from the above empirical example, the proposed approach retrieves solutions for a new online complaint according to its context (synthetically considering complaint products, problems, and content). It overcomes the limitation of keywords based retrieval and offers a personalized recommendation for online complainant handling at semantic level. Except the last step, all steps are completely automatic. Thus the approach provides a feasible way to reduce manual intervention in online complaint handling.

Figure 3. The details of the source case SC00000058.

5. Conclusion

This paper has proposed a similarity-based approach for retrieving similar cases accurately to deal with online complaints intelligently. This approach mainly consists of two parts: (1) The establishment of a global measure and (2) the determination of a global measure with high accuracy to assess the overall similarity of two online complaint cases. The first part comprehensively took into account the word-level, sense-level, and text-level features of online complaints and explained the process of using existing word-level, sense-level, and text-level measures to calculate the similarities between complaint products, between complaint problems, and between complaint contents. The second part described how to assign different weights and choose different local measures to obtain a global measure with the highest accuracy among the accuracies of all possible linear combinations of these chosen local measures for different case bases and how to use the obtained global measure to assess the overall similarity between online complaint cases. The paper has also reported the implementation and evaluation of the approach. The evaluation results suggest that an approach considering the features of similarity at word, sense, and text levels simultaneously can achieve higher accuracy than those approaches taking into account only one level feature of similarity and the determined global measure is more accurate than all of its linear combinations.

5.1. Practical implications

The paper has presented an intelligent online complaint handling approach based on case-based reasoning. This approach provides a feasible way to reduce manual intervention in online complaint handling. In the approach, a global similarity measure that linearly combines the similarity of complaint products, the similarity of complaint problems, and the similarity of complaint contents was established to accurately compute the similarity between online complaint cases. A managerial implication deriving from this establishment is that complaint products, problems, and content are all distinctive features of an online complaint. These features should be synthetically considered when handling an online complaint.

In addition, it can be found from the evaluation of the approach that a similarity measure simultaneously considering the features of similarity at word, sense, and text levels can obtain higher accuracy than those similarity measures only considering one level feature of similarity. A practical implication of this finding is that the presented determination method of a global similarity measure with high accuracy can be applied in other applications that consider multiple similarity features or linguistic levels.

5.2. Future research

Future research will aim especially at overcoming one major limitation of the proposed approach: The present paper mainly discussed how to improve the accuracy of similarity-based online complaint case retrieval. It did not consider the improvement of the speed of online complaint case retrieval. Like case retrieval accuracy, case retrieval speed is also a very important indicator to measure the performance of case retrieval. Hence, it is of significance to study how to improve online complaint case retrieval speed while case retrieval accuracy has been improved.

Acknowledgements

References

Chang, H.H., Tsai, Y.C., Wong, K.H., Wang, J.W. and Cho, F.J. (2015), "The effects of response strategies and severity of failure on consumer attribution with regard to negative word-of-mouth", Decis Support Syst, Vol. 71 No. 3, pp. 48–61.

Chen, C.K., Shie, A.J. and Yu, C.H. (2012), "A customer-oriented organisational diagnostic model based on data mining of customer-complaint databases", Expert Syst Appl, Vol. 39 No. 1, pp. 786–792.

Cheng, V.T. and Loi, M.K. (2014), "Handling Negative Online Customer Reviews: The Effects of Elaboration Likelihood Model and Distributive Justice", J Travel Tour Mark, Vol. 31 No. 1, pp. 1–15.

Cheung, C.M.K. and Thadani, D.R. (2012), "The impact of electronic word-of-mouth communication: A literature analysis and integrative model", Decis Support Syst, Vol. 54 No. 1, pp. 461–470.

Cho, Y., Im, I., Hiltz, R. and Fjermestad, J. (2002), "An analysis of online customer complaints: Implications for web complaint management", In Proc 35th Ann Hawaii Int Conf Syst Sci, Big Island, USA, 2002,

IEEE Computer Society, Washington, DC, USA, pp. 176–185.

Coussement, K. and Van den Poel, D. (2008), "Improving customer complaint management by automatic email classification using linguistic style features as predictors", Decis Support Syst, Vol. 44 No. 4, pp. 870–882.

De Mantaras, R.L., McSherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., Faltings, B., Maher, M.L., Cox, M.T., Forbus, K., Keane, M., Aamodt, A. and Watson, I. (2005), "Retrieval, reuse, revision and retention in case-based reasoning", Knowl Eng Rev, Vol. 20 No. 3, pp. 215–240.

Dong, Z. and Dong, Q. (2006), HowNet and the Computation of Meaning, World Scientific, River Edge, USA.

Etzioni, O. (2011), "Search needs a shake-up", Nature, Vol. 476 No. 7358, pp. 25–26.

Faed, A., Chang, E., Saberi, M., Hussain, O.K. and Azadeh, A. (2016), "Intelligent customer complaint handling utilising principal component and data envelopment analysis (PDA)", Appl Soft Comput, Vol. 47 No. 10, pp. 614–630.

Fellbaum C. (Ed.) (1998), WordNet: An Electronic Database, MIT Press, Cambridge, USA.

Fournier, S. and Avery, J. (2011), "The uninvited brand", Bus Horiz, Vol. 54 No. 3, pp. 193–207.

Galitsky, B.A., González, M.P. and Chesñevar, C.I. (2009), "A novel approach for classifying customer complaints through graphs similarities in argumentative dialogues", Decis Support Syst, Vol. 46 No. 3, pp. 717–729.

Gupta, N.K. (2013), "Extracting phrases describing problems with products and services from twitter messages", Computación y Sistemas, Vol. 17 No. 2, pp. 197–206.

Harrison-Walker, L.J. (2001), "E-complaining: A content analysis of an internet complaint forum", J Serv Mark, Vol. 15 No. 5, pp. 397–412.

Islam, A. and Inkpen, D. (2008), "Semantic text similarity using corpus-based word similarity and string similarity", ACM Trans Knowl Discov Data, Vol. 2 No. 2, pp. 10–25.

Jansen, B.J., Zhang, M., Sobel, K. and Chowdury, A. (2009), "Twitter power: Tweets as electronic word of mouth", J Am Soc Inf Sci Technol, Vol. 60 No. 11, pp. 2169–2188.

Lee, S.J. and Lee, Z. (2006), "An experimental study of online complaint management in the online feedback forum", J Organ Comp Electron Commer, Vol. 16 No. 1, 65–85.

Lee, C.H., Wang, Y.H. and Trappey, A.J. (2015), "Ontology-based reasoning for the intelligent handling of customer complaints", Comput Ind Eng, Vol. 84 No. 6, pp. 144–155.

Lin, D. (1998), "An information-theoretic definition of similarity", In Proc 15th Int Conf Machine Learning, Madison, USA, 1998, Morgan Kaufmann Publishers Inc., San Francisco, USA, pp. 296–304.

Lu, W., Qin, Y., Qi, Q., Zeng, W., Zhong, Y., Liu, X. and Jiang, X. (2016), "Selecting a semantic similarity measure for concepts in two different CAD model data ontologies", Adv Eng Inform, Vol. 30 No. 3, pp. 449–466.

Ma, H.N., Wu, J.N. and Dang, Y.Z. (2006), "Ontology-driven information retrieval system for customers' complaints of mobile communication services", In Proc 2006 Int Conf Manage Sci Eng, Lille, France, 2006, Harbin Institute of Technology Press, Harbin, China, pp. 30–35.

Mattila, A.S., Andreau, L., Hanks, L. and Kim, E.E. (2013), "The impact of cyberostracism on online complaint handling: Is "automatic reply" any better than "no reply"?", Int J Retail Distrib Manage, Vol. 41 No. 1, pp. 45–60.

McGuinness, D.L. and van Harmelen, F. (2004), "OWL Web Ontology Language Overview", available at: http://www.w3.org/TR/owl-features/ (accessed 28 September 2016).

Mihalcea, R., Corley, C. and Strapparava, C. (2006), "Corpus-based and knowledge-based measures of text semantic similarity", In Proc 21st National Conf Artif Intell, Boston, USA, 2006, AAAI Press, Palo Alto, USA, pp. 775–780.

Minaei-Bidgoli, B. and Akhondzadeh, E. (2010), "New approach of using association rule mining in customer complaint management", Int J Comput Sci Issues, Vol. 7 No. 5, pp. 142–147.

Pfeffer, J., Zorbach, T. and Carley, K.M. (2014), "Understanding online firestorms: Negative word-of-mouth dynamics in social media networks", J Mark Commun, Vol. 20 No. 1–2, pp. 117–128.

Pilehvar, M.T. and Navigli, R. (2015), "From senses to texts: An all-in-one graph-based approach for measuring semantic similarity", Artif Intell, Vol. 228 No. 11, pp. 95–128.

Sánchez, D., Batet, M., Isern, D. and Valls, A. (2012), "Ontology-based semantic similarity: A new feature-based approach", Expert Syst Appl, Vol. 39 No. 9, pp. 7718–7728.

Schwab, P.N. (2015), "Online complaint handling practices: Company strategies and their effects upon post complaint satisfaction", CEB Working Paper N°15/005, Université Libre de Bruxelles, Solvay Brussels School of Economics and Management, Centre Emile Bernheim, Brussels, Belgium, 10 February.

Sitko-Lutek, A., Chuancharoen, S., Sukpitikul, A. and Phusaval, K. (2010), "Applying social network analysis on customer complaint handling", Ind Manage Data Syst, Vol. 110 No. 9, pp. 1402–1419.

Sultan, A.B., Abidin, K.Z. and Abdullah, M.T. (2008), "The implementation of agent-based complaint management system", J Comput Sci, Vol. 8 No. 5, pp. 205–207.

Taieb, M.A.H., Aouicha, M.B. and Hamadou, A.B. (2014), "Ontology-based approach for measuring semantic similarity", Eng Appl Artif Intell, Vol. 36 No. 11, pp. 238–261.

Thomas, J.B., Peters, C.O., Howell, E.G. and Robbins, K. (2012), "Social media and negative word of mouth: Strategies for handing unexpecting comments", Atlantic Mark J, Vol. 1 No. 2, pp. 87–108.

Trappey, A.J.C., Lee, C.H., Chen, W.P., and Trappey, C.V. (2010), "A framework of customer complaint handling system", In Proc 7th Int Conf Service Syst Service Manage, Tokyo, Japan, 2010, IEEE Computer Society, Washington, DC, USA, pp. 1–6.

Turney, P. (2001), "Mining the web for synonyms: PMI-IR versus LSA on TOEFL", In Proc 12th European Conf Machine Learning, Freiburg, Germany, 2001, Springer Berlin Heidelberg, Berlin, Germany, pp. 491–502.

Van Noort, G. and Willemsen, L.M. (2012), "Online damage control: the effects of proactive versus reactive webcare interventions in consumer-generated and brand-generated platforms", J Interact Mark, Vol. 26 No. 3, pp. 131–140.

Wu, Z. and Palmer, M. (1994), "Verbs semantics and lexical selection", In Proc of the 32nd Annual Meeting on Association for Computational Linguistics, Las Cruces, New Mexico, 1994, Association for Computational Linguistics, Stroudsburg, USA, pp. 133–138.

Zhang, X., Hou, X., Chen, X. and Zhuang, T. (2013), "Ontology-based semantic retrieval for engineering domain knowledge", Neurocomputing, Vol. 116 No. 9, pp. 382–391.

Zirtiloğlu, H. and Yolum, P. (2008), "Ranking semantic information for e-government: Complaints management", In Proc 1st Int Workshop Ontology-Supported Bus Intell, Karlsruhe, Germany, 2008, ACM, New York, USA, pp. 1–7.