A Knowledge-rich Similarity Measure for Improving IT Incident Resolution Process

Yong-Bin Kang, Arkady Zaslavsky,
Shonali Krishnaswamy
Caulfield School of IT
Monash University, Australia
{yongbin.kang, arkady.zaslavsky,
shonali.krishnaswamy}@infotech.monash.edu.au

Claudio Bartolini
HP Labs
Palo Alto, USA
claudio.bartolini@hp.com

ABSTRACT

The aim of incident management is to restore a given IT service disruption, simply called incident, to normal state as quickly as possible. In incident management, it is essential to resolve a new incident efficiently and accurately. However, typically, incident resolution process is largely manual, thus, it is time-consuming and error-prone. This paper proposes a new knowledge-rich similarity measure for improving this process. The role of this measure is to retrieve the most similar past incident cases for a new incident without human intervention. The solution information contained the retrieved incident cases can be utilized to resolve the new incident. The main feature of our similarity measure is to incorporate additional useful meta knowledge, outside of incident description that is the only exploited information in typical similarity measures used in CBR, to improve effectiveness. Moreover, this measure exploits as much semantic knowledge as possible about features contained in previous incident cases. Through an experimental evaluation, we show the effectiveness, technical coherence and feasibility of this measure using a real dataset.

Categories and Subject Descriptors

H.4.2 [**Information Systems Applications**]: Types of Systems—*Decision support*

Keywords

Knowledge-rich similarity measure, IT Service Management, IT incident management, Incident resolution process

1. INTRODUCTION

The main objective of IT Service Management (ITSM) is to advance IT best practices in service delivery and service support [13]. IT Infrastructure Library (ITIL [2]) has been recognized as the world de facto standard, which provides a comprehensive set of principles advising on best ITSM practice. In general, *incident management* is the most important element of ITIL process model for delivering IT services [8]. ITIL defines an *incident* as any event which is not part of the standard operation of a service and which causes an interruption to the quality of that service. The goal of incident management is to restore normal service operation as quickly as possible and minimize the adverse impact on business operations. A typical IT support organization is structured as a complex network of *workgroups*, each comprising of a set of skilled *operators* [4]. Commonly, workgroups are divided into a few support levels (usually 3 - 5), where higher level workgroups are more specialized, dealing with more difficult and time-consuming issues. Meanwhile, the help-desk is assigned to level 0 and provides a frontier interface for customers reporting an IT disruption.

A typical incident management is processed as follows [4, 13]: (1) *incident detection*: given an IT disruption reported by a customer, the help-desk creates a new incident, describing the symptom or customer's perception of the disruption. This description is often called as *incident description*; (2) *incident classification*: the help-desk estimates the classification of the incident, which will be used to support initial incident resolution and help to escalate the incident to the appropriate workgroup; (3) *initial incident resolution*: initially, the help-desk guesses possible keywords of the incident based on its incident description along with his/her intelligence and experience. With the help of the keywords and the estimated incident classification, the help-desk attempts search procedures to retrieve the most similar incident cases stored in a database for the incident by *manual* work. If any matched incident is found, its solution may help to resolve the incident, otherwise go to the next step; (4) *incident escalation and resolution*: the incident is escalated to the appropriate workgroup at a higher level to be resolved with more specialized skill. Here, incident resolution process is analogously carried out as process (3). Besides, the incident information may be updated, if necessary, by the assigned workgroup. If the incident is resolved, it is closed, otherwise, repeatedly escalated to other workgroups until resolved.

Considering the typical incident management above, we can easily notice that an incident I can contain various information about a customer reporting I, I's incident description, I's incident classification, and a workgroup responsible for resolving I.

However, the typical incident management has two prob-

SAC'10 March 22-26, 2010, Sierre, Switzerland.
Copyright 2010 ACM 978-1-60558-638-0/10/03 ...$10.00.

lems: First, "incident resolution" process is largely manual, thus, it is time-consuming and error-prone. Second, we cannot guarantee that similar repetitive incidents are resolved in *consistent* manners. This is because there is the high possibility of generating different resolutions for such repetitive incidents, depending on which workgroups are assigned to handle them. For instance, given an incident, a lower level workgroup may produce too many candidate solutions, thus choose a wrong final resolution from them. Meanwhile, a higher level workgroup is more skillful, thereby producing a more accurate resolution. This problem would be more serious in a situation where the positions of workgroups are frequently changed in IT organizations today.

To address the above problems, this paper proposes a new knowledge-rich similarity measure for improving the typical "incident resolution" process. The main role of this measure is to retrieve the most similar incident cases for a given incident effectively without human (i.e., help-desk or workgroups) intervention. The distinctive feature of this measure is to incorporate the following two additional useful meta knowledge of incident into its computational space, beside incident description : *incident classification* and *workgroup*. Moreover, this similarity measure can exploit semantic domain knowledge about various information contained in the contents of incident cases.

This paper is organized as follows: In Section 2, we review related work. In section 3, we introduce our computer-facilitated incident management with its core components. Then, we present a new knowledge-rich similarity measure in Section 4. An experimental evaluation is presented in Section 5, and then we conclude this paper in Section 6.

2. RELATED WORK

To accurately retrieve the most similar incident cases for a given incident, intelligent systems are needed to provide useful decision support for the retrieval. A practical approach that has been adopted in these systems is *Case-Based Reasoning (CBR)*[23]. Here, we present the related work of similarity measurement in the domain of CBR.

The main principle of CBR is that an optimal solution for a given problem is retrieved on the basis of similar experience learned in the past [23]. A unit of the experience is represented as a *case* stored in a *case base*. A case contains a combination of useful information to solve a given problem, such as problem description and solution description. The main premise underlying CBR is that the more similar two problem descriptions are, the more similar their solutions are. Thus, in CBR, retrieving an appropriate solution for a given problem is mostly relied on a certain similarity measure between the problem description of a request (e.g., incident) and the problem description of a case. However, one main limitation of similarity measures used in CBR lies in their very limited computational space. That is, these similarity measures are computed within a single computational space, i.e., the problem description space of a request and a case being compared, ignoring a set of additional useful information available.

For example, the common principle of the similarity measures used in [5, 6, 24, 25] is described as follows: once a request is presented, the system tries to retrieve similar cases to the request. First, problem features (or description) are extracted from the request, and the retrieval is performed based on a similarity computation between the problem features and cases. The weights of the problem features are specified by the user or system, and the similarity computation is decided by the amount of information values between the problem description and cases (e.g., the weight of the problem features in the candidate cases).

To extend the computational space for similarity measurement, diverse approaches have been proposed by attempting to utilize various types of knowledge. For instance, HOMER [9] and IHDF [12] incorporate *decision rules* to retrieve better solution cases. In the systems, closely related cases are first retrieved by computing similarities between two descriptions of a given problem and past cases, and then the decision rules are applied to further narrow them down. The decision rules are represented using a set of declarative and procedural types of knowledge (a set of the pairs of problem descriptions and their corresponding solutions). Although, these systems try to extend a computational space for similarity measurement using decision rules, however, such rules are mainly derived from information units only drawn from the problem descriptions of cases.

More recently, various forms of knowledge-intensive CBR systems have been proposed, which emphasize the importance of exploiting richer knowledge about the application domain[3, 20, 21]. With the advancement of ontology research, these systems are mainly focusing on similarity measurement taking the advantages of *ontological* description to model the domain knowledge. The main aim of using this description is to characterize a conceptualization of the most important aspects of domain knowledge by describing domain-specific concepts and their relationships in a formal way. By using ontological description, these systems can identify implicit semantic knowledge about these relationships that may assist to make more complete similarity measurement. Although, these systems adopt the knowledge-intensive similarity measures based on ontological description about domain knowledge, this description is only used to enhance the semantic meaning of a representation of the problem features (description) of cases.

3. COMPUTER-FACILITATED INCIDENT MANAGEMENT

The proposed conceptual model for incident management is depicted in Fig. 1. The distinctive feature of this model from the typical incident management is that incident resolution procedure can be performed, based on a knowledge-rich similarity measure. Given an IT disruption, the help-desk creates an incident and then the incident resolution procedure is taken to retrieve the most k-top similar incident cases for the incident. Once these incident cases are returned to the help-desk or workgroups, the help-desk may describe a solution for the incident on the basis of them. If the solution is satisfied, the incident is closed. Otherwise, it is escalated to an appropriate higher workgroup. The assigned workgroup at a higher level reviews the forwarded incident and the incident information can be updated or more specialized, if necessary. Then, the incident resolution procedure is performed again. This process is repeatedly continued until the incident is solved.

In the following subsections, we describe the components in Fig. 1 and then present the proposed knowledge-intensive similarity measurement in Section 4.

3.1 Incident Classification Information

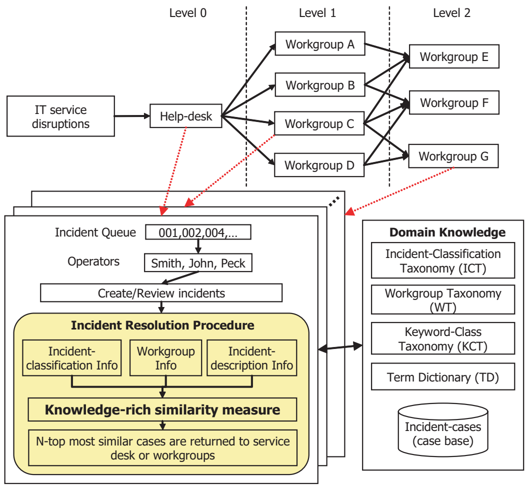

Figure 1: The proposed conceptual model for incident management.

Incident classification information, specified by the help-desk or workgroups, provides useful support for finding the most similar incident cases for a given incident. Typically, the entries of incident classification have hierarchical relationships, such as "`fault - software - database`". In order to describe these entries and their relationships, we model a specific taxonomy, called *Incident-Classification Taxonomy (ICT)*, using "is-a" relationship [16]. An example of ICT is shown in Fig. 2(a). The semantic knowledge about entries existed in ICT will facilitate comparative analysis in our similarity measure.

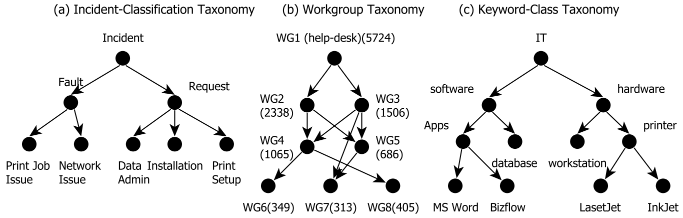

Figure 2: An example of three taxonomies.

3.2 Workgroup Information

In the incident resolution procedure in Fig. 1, workgroups also provide an effective coordination of activities for solving given incidents. Such activities include creating, reviewing, and updating the contents of the incidents. In our work, workgroup information is represented by workgroup *level* information. To describe the workgroup level information, we model a taxonomy, named *Workgroup Taxonomy (WT)* (see Fig. 2(b)). Since workgroups are mostly divided into a few support levels, the semantic knowledge about their relationships provided by WT will be used as useful information in our similarity measure.

3.3 Incident Description Information

Incident description information represents meaningful terms which are possibly extracted from incident description. These terms can be used as useful criteria to retrieve the most similar incident cases for a given incident. An issue raised here is how to obtain these terms. To address this issue, we present three types of terms:

Table 1: Term Dictionary (TD) and incident description information.

Standardized-Term	Alias-Terms	Keyword-Class
print	print, printer, printing error, etc	printer
w3	w3, world wide web, www, etc	internet
Bizflow	bizflow, bizflow program, etc	Bizflow

An incident description: "printer tray cannot print from Bizflow."
-tokens: printer, tray, unable, print, Bizflow.
-keywords: print(← printer), print, Bizflow
-keyword-classes: printer, printer, Bizflow

Tokens: Given an incident description, tokens are obtained by separating it into terms, removing stop-words (e.g, punctuation, 'a', 'is', 'the', etc), and stemming the terms; **Keyword:** Given tokens, keywords are extracted by selecting domain-specific standardized terms, exploiting *Term Dictionary* (TD). TD maintains standardized domain-specific terms and their alias terms used in a given IT domain. The structure of TD consists of a tuple of `Standardized-Term`, `Alias-Terms`, and `Keyword-Class`. The `Keyword-Class` is fractionalized in detail (e.g., internet, database, printer, etc) on the basis of classifications in IT glossaries[1]. The main reason for using TD is that if keywords are always described using a set of standardized domain-specific terms, we can avoid the ambiguity occurred from these causes: some heterogeneous keywords can be used to describe a same keyword, and alias keywords often cannot be interpreted properly or missed; **Keyword-classes:** Keyword-classes represent the fractionalized keyword-classes, listed under the column of the `Keyword-Class` of TD. Further, *Keyword-Class Taxonomy* (KCT) describes these keyword-classes and their relationships using "is-a" ontology (see Fig. 2(c)). The aim is to improve the possibility of capturing implicit semantic knowledge about the keyword-classes using this ontological description. Table 1 shows an example of TD and how tokens, keywords, and keyword-classes are extracted from an incident description.

4. KNOWLEDGE-RICH SIMILARITY MEASURE

In this section, we present a new knowledge-rich similarity measure. This measure is used to retrieve the most similar k-top incident cases for a given incident. The main feature of this measure lies in the incorporation of the following knowledge types:

(1) Incident description with two additional meta knowledge: No doubt, incident description is the most fundamental information to retrieve the most similar incident cases for a given incident. Besides, incident classification information may also provide useful support for retrieving them. In the typical incident resolution procedure, some benefits of using this information are considered as prop-

[1] To make a relevant set of IT glossaries, we referred to a number of terminologies used in ITSM and IT domains from these websites:
http://www.covestic.com/downloads/glossary_itsm_terms.pdf,
http://www.itsmf.ca/about/glossary.html,
http://www.e-help-desk.com/glossary.html,
http://www.cryer.co.uk/glossary,
http://www.itcom.itd.umich.edu/glossary.html

erly routing incidents to the correct workgroup, speeding up diagnoses by collecting the right information, and improving the efficiency of workgroups. Motivated by the benefits, we exploit this information into our computational space of similarity measurement. Moreover, we exploit workgroup information into the space as well. Workgroups have main responsibilities for the activities of specifying incident classification and describing incident description as mentioned in Section 3. However, such activities can be achieved in various ways, even if incoming IT disruptions are similar to each other, according to different workgroups handling them. Thus, given an incident, we believe that if we utilize its incident classification and incident description information into the computational space of similarity measurement, we also need to exploit workgroup information dealing with it into this computational space.

(2) Three different types of semantic domain knowledge: In order to take the advantages of using taxonomic knowledge, we exploit three taxonomies, ICT, WT, and KCT in our computational space of similarity measurement. Exploiting the relationships among the entries in such taxonomies will contribute to making more complete similarity measurement.

4.1 Modeling Similarity Measure

To measure a semantic similarity between a given incident and an incident case, we define an object model for them. For this, we choose the set-based model[7] due to its simplicity. In this model, objects are treated as an ordered set of three types of information: incident classification, workgroup, and incident description - e.g., {'`Print Job Issue`', '`WG1`', '`Printer is unable to print Bizflow`'}.

The goal of our similarity measure is to compute similarity $\mathtt{SIM}(O_1, O_2)$, a real number [0,1], where O_1 and O_2 are given two objects. Given incident object $O_1 = \{a_{c1}, a_{w1}, a_{d1}\}$ and incident case object $O_2 = \{a_{c2}, a_{w2}, a_{d2}\}$, their similarity is defined to be

$$\begin{aligned}\mathtt{SIM}(O_1, O_2) =& w_1 * sim_c(a_{c1}, a_{c2}) + \\ & w_2 * sim_w(a_{w1}, a_{w2}) + w_3 * sim_d(a_{d1}, a_{d2})\end{aligned} \quad (1)$$

where $sim_c(a_{c1}, a_{c2})$ denotes similarity measure between two incident classifications (a_{c1}, a_{c2}), $sim_w(a_{w1}, a_{w2})$ means similarity measure between two workgroups (a_{w1}, a_{w2}), and sim_d (a_{d1}, a_{d2}) represents similarity measure between two incident descriptions (a_{d1}, a_{d2}). To express the different importance of these three measures, we define weights on them, w_1, w_2, and w_3, respectively $(w_1 + w_2 + w_3 = 1.0)$. Now, our main concern is how to define these three similarity measures, exploiting semantic knowledge described in taxonomies ICT, WT and KCT.

4.2 Similarity Measure for Incident Classifications

Computing semantic similarity between two entities resided in a taxonomy can be generally classified into two approaches [15]: *distance-based* approach calculating distance between them in the taxonomy, and *node-based* approach estimating the amount of sharing *information content* between two entities. In this approach, the more information content they have, the more similar they are.

For the node-based approach, there are two methods of quantifying the information content of concept c using a given taxonomy. First, Resnik [17] quantifies it by considering the negative log likelihood as denoted by $ic_{res}(c) = -\log p(c)$ $(p(c) = \frac{freq(c)}{N})$, where $freq(c)$ is the sum of the frequency counts of its all subsumed concepts occurred in a given corpus, and N is the total number of concepts in the taxonomy. Second, Seco [19] exploits only structural information of the taxonomy, not relying on corpora analysis. This method is denoted as $ic_{se}(c) = 1 - \frac{\log(|sc(c)|+1)}{\log(|T|)}$, where $sc(c)$ the set of subsumed concepts of c and $|T|$ is the total number of concepts belonging to taxonomy T. The denominator equivalent to the value of the most informative concept, serves as a normalizing factor assuring that $ic_{se}(c)$ are in [0,1].

To compute the similarity $sim_c(a_{c1}, a_{c2})$ (see eq.1), we combine Jiang's measure [11] with Seco's method. The reason is that several studies commonly found that Jiang's measure is the most effective and outperforms other approaches [11, 15, 19]. Moreover, since we do not need a corpus, we choose $ic_{se}(c)$ to measure the information content of incident classification c. Consequently,

$$sim_c(a_{c1}, a_{c2}) = 1 - \Big(\frac{ic_{se}(a_{c1}) + ic_{se}(a_{c2}) - 2 * ic_{se}(lcs(a_{c1}, a_{c2}))}{2}\Big), \quad (2)$$

where $lcs(a_{c1}, a_{c2})$ is the least common subsumer that subsumes a_{c1} and a_{c2} in a given taxonomy ICT. This formulation guarantees that this similarity score is normalized between interval [0,1]. As an example, considering the taxonomy ICT in Fig. 2(a), let a_{c1} = '`Print Job Issue`' and a_{c2}= '`Network Issue`', $sim_c(a_{c1}, a_{c2}) = 1 - \big(\frac{1.0+1.0-2*.472}{2}\big) =$ .472, where $lcs(a_{c1}, a_{c2})$ = '`Fault`'.

4.3 Similarity Measure for Workgroups

Given two workgroups a_{w1} and a_{w2}, $sim_w(a_{w1}, a_{w2})$ is computed by exploiting their hierarchical relationship inherent in a given taxonomy WT. To define this similarity, we utilize the concept of the node-based approach. As described in Section 4.2, to compute a similarity based on the node-based approach, we need to quantify the information content of every workgroup in the WT.

The information content of workgroup a_w, denoted as $ic_w(a_w)$, is estimated by combining the statistical information of the incident cases handled by a_w and the depth information of a_w in the WT. Based on Resnik [17], $ic_w(a_w)$ can be quantified by considering the negative log likelihood. That is,

$$ic_w(a_w) = -\log p(a_w), \quad (3)$$

where $p(a_w) = \frac{freq(a_w)}{N}$, where $freq(a_w)$ is the sum of the incident cases handled by a_w's all subsumed workgroups existed in the WT, and N is the total number of incident cases stored in a case base. Intuitively, as $p(a_w)$ increases, a_w becomes more abstract and the informativeness of a_w decreases. For instance, referring to the WT of Fig. 2(b), the numbers in the parentheses denote the numbers of incident cases handled by workgroups $\mathtt{WG}_1$ - $\mathtt{WG}_8$. Using eq.3, $ic_w(\mathtt{WG}_2)$ = .381 and $ic_w(\mathtt{WG}_4)$ = .643.

Therefore, the similarity $sim_w(a_{w1}, a_{w2})$ can be computed by combining Jiang's measure with eq.3. Then,

$$sim_w(a_{w1}, a_{w2}) = 1 - \Big(\frac{ic_w(a_{w1}) + ic_w(a_{w2}) - 2 * ic_w(lcs(a_{w1}, a_{w2}))}{2}\Big), \quad (4)$$

where $lcs(a_{w1}, a_{w2})$ is the least common subsumer of a_{w1} and a_{w2} in the WT. This similarity is normalized to real numbers between [0,1].

4.4 Similarity Measure for Incident Descriptions

Given two incident descriptions a_{d1} and a_{d2}, $sim_d(a_{d1}, a_{d2})$ is computed by combining the following three measures **M1**, **M2**, and **M3**. These measures exploit tokens, keywords, and keyword-classes, respectively, obtained after the text processing described in Section 3.3. Besides, all these measures are normalized between 0 (completely dissimilar) and 1 (identical), becoming to be 1 as the compared entities have more and more commonality.

(M1) Similarity between tokens: Given two sets of tokens t_{d1} and t_{d2}, their similarity is computed by extending the typical intersection-based similarity measures, such as Jaccard's and Dice's coefficients [22]. In these measures, the *commonality*, denoted as '$\cap$', between two tokens is defined as: if two tokens are identical, then their commonality is 1, otherwise 0. But the limitation of the measures using this commonality is that if $t_{d1} \cap t_{d2} = \emptyset$, then their similarity is 0. This may be unreasonable in a situation where two tokens have a similar *semantic* meaning. In this situation, it would be more desirable if these two sets have nonzero similarity even though they have no identical tokens in common.

To overcome this limitation, we extend the concept of the commonality by using semantic knowledge about tokens inherent in WordNet[2][14]. With using WordNet, we define an extended concept of the commonality, denoted as '$\cap^+$', to be: if two tokens are identical, or are in the same synonym sets of WordNet, then their commonality is 1, otherwise 0.

Given t_{d1} and t_{d2}, the proposed similarities between them using the concept of '$\cap^+$', denoted by sim_{jc^+} and sim_{dc^+}, are defined by $sim_{jc^+}(t_{d1}, t_{d2}) = \frac{|t_{d1} \cap^+ t_{d2}|}{|t_{d1} \cup t_{d2}|}$ and $sim_{dc^+}(t_{d1}, t_{d2}) = \frac{2 * |t_{d1} \cap^+ t_{d2}|}{|t_{d1}| + |t_{d2}|}$, respectively.

(M2) Similarity between keywords: Given two sets of keywords k_{d1} and k_{d2}, we compute their similarity using the traditional set-based similarity measures such as Jaccard's and Dice's coefficients. This similarity measure is based on the premise that the more keywords k_{d1} and k_{d2} have in common, the more similar their incident descriptions are. Given k_{d1} and k_{d2}, their Jaccard's and Dice's coefficients, denoted as sim_{jc} and sim_{dc}, are represented by $sim_{jc}(k_{d1}, k_{d2}) = \frac{|k_{d1} \cap k_{d2}|}{|k_{d1} \cup k_{d2}|}$ and $sim_{dc}(k_{d1}, k_{d2}) = \frac{2 * |k_{d1} \cap k_{d2}|}{|k_{d1}| + |k_{d2}|}$, respectively.

(M3) Similarity between keyword-classes: Given two sets of keyword-classes kc_{d1} and kc_{d2}, we compute a semantic similarity of them, exploiting semantic knowledge resided in a given taxonomy KCT. The premise is that given two keywords, if their keyword-classes are similar, these kyewords are also similar semantically. Given kc_{d1} and kc_{d2}, recall that the limitation of the typical intersection-based similarity measures is: if $kc_{d1} \cap kc_{d2} = \emptyset$, then their similarity is zero. However, this may be also unreasonable in a situation where two keywords are siblings in the KCT. In the situation, these two sets of keywords should have nonzero similarity even though they have no identical keywords in common. The inclusion of semantic knowledge inherent in the KCT can allow the intersection-based similarity to avoid such inaccurate zero similarity measure.

To compute a similarity between kc_{d1} and kc_{d2}, we extend the typical intersection-based similarity measures by combining the node-based approach. Recall that the idea of the node-based approach is that the similarity of two entities belonging to a given taxonomy is defined by the information content of their least concept subsumer (LCS). Based on the concept of LCS, we augment $kc_{d1} = \{kc_{11}, ..., kc_{n1}\}$ and $kc_{d2} = \{kc_{12}, ..., kc_{m2}\}$ by including the set of the LCSs that subsume all pairs of $kc_{i1(1 \leq i \leq n)}$ and $kc_{j2(1 \leq j \leq m)}$. Given kc_{d1} and kc_{d2}, their *augmented* keyword-classes, which denote kc^+_{d1} and kc^+_{d2}, can be defined to be $kc^+_{d1} = kc_{d1} \cup \{lcs(kc_{i1}, kc_{j2})\}$ and $kc^+_{d2} = kc_{d2} \cup \{lcs(kc_{i1}, kc_{j2})\}$, where $lcs(kc_{i1}, kc_{j2})$ is the LCS of keyword-classes kc_{i1} and kc_{j2}.

Based on the augmented sets of keyword-classes, we then assign *weight* to every keyword-class in kc^+_{d1} and kc^+_{d2}. For this, we define this weight to be the information content of itself using Seco's method (see Section 4.2). That is, let $w(kc)$ be the weight of keyword-class kc, then $w(kc) = ic_{se}(kc)$. Formally, the proposed similarities between kc_{d1} and kc_{d2} using the concept of the kc^+_{d1} and kc^+_{d2}, denoted by sim_{jc^T} and sim_{dc^T}, respectively, are defined by

$$sim_{jc^T}(kc_{d1}, kc_{d2}) = \frac{\sum_{\{i|kc_i \in kc^+_{d1} \cap kc^+_{d2}\}} w(kc_i)}{\sum_{\{j|kc_j \in kc^+_{d1} \cup kc^+_{d2}\}} w(kc_j)},$$

$$sim_{dc^T}(kc_{d1}, kc_{d2}) = \frac{2 * \sum_{\{i|kc_i \in kc^+_{d1} \cap kc^+_{d2}\}} w(kc_i)}{\sum_{\{j|kc_j \in kc^+_{d1}\}} w(kc_j) + \sum_{\{j|kc_j \in kc^+_{d2}\}} w(kc_j)}.$$

To sum up, given two incident descriptions a_{d1} and a_{d2}, the similarity $sim_d(a_{d1}, a_{d2})$ is computed by combining the above three similarity measures. For example, using the concept of Jaccard's coefficient, $sim_d(a_{d1}, a_{d2})$ can be represented by

$$\begin{aligned} sim_d(a_{d1}, a_{d2}) = {} & \alpha * sim_{jc^+}(t_{d1}, t_{d2}) + \\ & \beta * sim_{jc}(k_{d1}, k_{d2}) + \gamma * sim_{jc^T}(kc_{d1}, kc_{d2}) \end{aligned} \tag{5}$$

where α, β, and γ are combination coefficients representing acceptable reliability degrees of the corresponding three similarities, where $\alpha + \beta + \gamma = 1.0$.

5. EVALUATION

Our evaluation goal is to demonstrate the effectiveness of our similarity measurement. The evaluation is performed using a real dataset based on off-line analysis. The effectiveness of our approach is defined by the well-known metrics of *precision* and *recall*.

In our evaluation, an empirical study was applied to determine whether two information components (i.e., incident classification and workgroup) are individually useful, and whether our knowledge-rich similarity measure itself outperforms a similarity measure using only incident description information as used in the typical similarity measures in CBR.

For this study, we are interested in comparing four similarity measures using the following combinations: (1) only incident description ($\texttt{SIM}_{\rm M1}$), (2) workgroup and incident description ($\texttt{SIM}_{\rm M2}$), (3) incident classification and incident description ($\texttt{SIM}_{\rm M3}$), and (4) all of incident classification, workgroup, and incident description ($\texttt{SIM}_{\rm M4}$). Formally, given incident object $O_1 = \{a_{c1}, a_{w1}, a_{d1}\}$ and incident case object $O_2 = \{a_{c2}, a_{w2}, a_{d2}\}$, these measures are represented as

$$\begin{aligned} &\texttt{SIM}_{\rm M1}(O_1, O_2) = sim_d(a_{d1}, a_{d2}), \\ &\texttt{SIM}_{\rm M2}(O_1, O_2) = c * sim_w(a_{w1}, a_{w2}) + (1 - c) * sim_d(a_{d1}, a_{d2}), \\ &\texttt{SIM}_{\rm M3}(O_1, O_2) = c' * sim_c(a_{c1}, a_{c2}) + (1 - c') * sim_d(a_{d1}, a_{d2}), \\ &\texttt{SIM}_{\rm M4}(O_1, O_2) = \texttt{SIM}(O_1, O_2)(\text{eq.1}), \end{aligned}$$

[2] WordNet is a broad coverage lexical network of English words being organized into networks of synonym sets (synsets), which represents semantic meanings of English words.

Table 2: 9 test incident cases and query incident.

ID	Work group	Incident classification	Incident description	Solution description
1	WG5	Fault	Bizflow running slow again	Link in Orpington has been upgraded and an extra server has been installed into the Bizflow farm.
2	WG3	Network Issue	Bizflow running very slowly	There is a new server being set up in Orpington to solve this issue. 20 users so far have access to it as of 2nd October. She said that the users on the new server have seen a big improvement in the period.
3	WG2	Print Setup	Break-Fix user is unable to print from Bizflow	Confirmed with Lucy printing OK now.
4	WG3	Fault	Bizflow printing error	Have assisted the user and her colleagues with printing issue and tested all working fine. Think the spooler service need to be restarted on the citrix server
5	WG8	Print Job Issue	unable to print from bizflow to the printer uktr7940	Richard Walker : no problem found
6	WG5	Fault	Bizflow printing issue	Restarted print spooler on Citrix42
7	WG3	Fault	User cannot find printer	User can print but can't select the tray. User is trying to print from different Trays. Printer don't recognise the Tray she selected
8	WG2	Print Job Issue	User cannot print from any printer	Lucy confirmed all printing OK now. TW
9	WG5	Print Job Issue	Printing problem	This was caused by network and server errors yesterday. These services have been restored.
Q	WG5	Network Issue	Unable to print from Bizflow	Restarted print spooler on citrix53.

Table 3: Test result using the dataset in Table 2.

Measures	ID	Similarity Score	$sim_d(a_{d1}, a_{d2})$	$sim_w(a_{w1}, a_{w2})$	$sim_c(a_{c1}, a_{c2})$
$\texttt{SIM}_{\mathrm{M1}}$	3	.800	.800	N/A	N/A
	5	.800	.800	N/A	N/A
	4	.750	.750	N/A	N/A
	6	.750	.750	N/A	N/A
	2	.368	.368	N/A	N/A
$\texttt{SIM}_{\mathrm{M2}}$	6	.800	.750	1.0	N/A
	3	.769	.800	.644	N/A
	4	.736	.750	.682	N/A
	5	.658	.800	.091	N/A
	2	.494	.368	1.0	N/A
$\texttt{SIM}_{\mathrm{M3}}$	5	.745	.800	N/A	.524
	4	.721	.750	N/A	.607
	6	.721	.750	N/A	.607
	3	.640	.800	N/A	0.0
	2	.494	.368	N/A	1.0
$\texttt{SIM}_{\mathrm{M4}}$	6	.761	.750	1.0	.607
	4	.729	.750	.682	.607
	3	.704	.800	.644	0.0
	5	.702	.800	.091	.524
	2	.494	.368	1.0	1.0

where $sim_d(a_{d1}, a_{d2})$, $sim_w(a_{w1}, a_{w2})$ and $sim_c(a_{c1}, a_{c2})$ are defined by eq.5, eq.4 and eq.2, respectively, and c and c' are combination coefficients.

Our test collection consists of three components: (1) Incident case base (CB): a real-life incident management dataset (12386 incident cases) was gathered from an installation of HP Service Manager [1]. In the CB, the solutions are described by workgroups' manual work, based on the typical incident management procedure, (2) TD: TD was constructed using the IT glossaries (see footnote[1]), (3) Taxonomies ICT, WT, KCT: ICT was constructed using 51 incident classifications existed in the CB. WT was built by observing the associations of the 8 different workgroups $\texttt{WG}_1$ - $\texttt{WG}_8$ within the CB (see Fig. 2(b)). The associations were identified by the statistical information of the incident cases handled by them. Regarding KCT, although its complete definition may be impossible, we defined a KCT based on the TD (33 keyword-classes).

5.1 Demonstration

To help understand how $\texttt{SIM}_{\mathrm{M1}}$ - $\texttt{SIM}_{\mathrm{M4}}$ are computed, we begin with a simple demonstration. Let us consider 9 incident cases chosen from the CB, as seen in Table 2. Among them, 7 are related to '`printing problem`' and the remaining 2 are linked with '`Bizflow problem`', but all 9 incidents have some common tokens and keywords. As a query incident, we used a real incident case chosen from the CB (see `ID`='Q' in Table 2). Let Q be the query incident and $C = \{O_1, ..., O_9\}$ be the collection of the 9 incident cases, where $O_{k(1\leq k\leq 9)}$ is k^{th} incident case whose `ID` is k.

Here, our goal is to retrieve the most similar incident cases for Q with regard to *solution description*. Referring to Table 2, we can easily identify that Q' solution is about '`restarting printer spooler`'. Among the incident cases in C, the ones having the most similar solutions with Q are O_4 and O_6. Therefore, the effectiveness can be determined by looking at how O_4 and O_6 are highly ranked in $\texttt{SIM}_{\mathrm{M1}}$ - $\texttt{SIM}_{\mathrm{M4}}$.

In Table 3, using C and Q, the result of $\texttt{SIM}_{\mathrm{M1}}$ - $\texttt{SIM}_{\mathrm{M4}}$ is shown, ranking 5-top incident cases by means of similarity score. First, $\texttt{SIM}_{\mathrm{M1}}$ ranked O_3 and O_5 as 2-top incident cases. To compute $\texttt{SIM}_{\mathrm{M1}}$, we assumed that the combination coefficients α, β and γ (see eq.5) are simply equal to each other, i.e., α=β=γ=1/3. These coefficients were identically assigned to all $\texttt{SIM}_{\mathrm{M1}}$ - $\texttt{SIM}_{\mathrm{M4}}$. Second, $\texttt{SIM}_{\mathrm{M2}}$ ranked O_6 and O_3 as 1^{th} and 2^{th} incident cases, respectively, while O_5 ranked as 4^{th} incident case, since sim_w("$\texttt{WG}_5$", "$\texttt{WG}_8$") = .091 (relatively very low). To determine the combination coefficient c for $\texttt{SIM}_{\mathrm{M2}}$, we run $\texttt{SIM}_{\mathrm{M2}}$ with various values of c on the CB. Through this examination, we set c = .2 as an optimal coefficient. Third, to measure $\texttt{SIM}_{\mathrm{M3}}$, we also needed to set the combination coefficient c'. The c' was set to be also .2 after taking the same examination as for determining c. This measure ranked O_5 and O_4 as 1^{th} and 2^{th} incident cases, respectively, while O_3 (1^{th} incident case in $\texttt{SIM}_{\mathrm{M1}}$) ranked as 4^{th} incident case, because sim_c("`Network Issue`","`Print Setup`")=0.0. Finally, to compute $\texttt{SIM}_{\mathrm{M4}}$, we set the weights w_1 = .1, w_2 = .1 and w_3 = .8 for eq.1. The reason is that when considering two coefficient values, c = .2 and c' = .2, we would not expect that the importance of $sim_d(a_{d1}, a_{d2})$ cannot be less than .8 of 1.0, and also we assumed that the importance of w_1 and w_2 are equivalent to each other. This measure ranked O_6 and O_4 as 2-top incident cases, meeting our goal in this experiment. As a result, we found that our approach $\texttt{SIM}_{\mathrm{M4}}$ is the most effective among $\texttt{SIM}_{\mathrm{M1}}$ - $\texttt{SIM}_{\mathrm{M4}}$.

5.2 Measuring the Effectiveness

In our evaluation, the measures of precision (`P`) and recall (`R`) focus an empirical evaluation on the returns of *relevant* and *irrelevant* incident cases. To measure `PR` in the standard way, we need two things: (1) A test suite of information needs (queries): our test suite consists of 3 *query-sets* of incident cases which were randomly selected from the CB, each comprising of 10 incident cases (called *query incidents*). (2) A set of relevance judgments: a set of binary assessments of either relevant or irrelevant for all pairs of query incidents and the corresponding incident cases retrieved.

To measure `PR`, we carried out 3 assessments by 3 PhD students who have strong background of IT. In each assessment, one of the query-sets (10 query incidents) was used and for this query-set relevance judgments over k-top (k=10) incident cases returned by $\texttt{SIM}_{\mathrm{M1}}$ - $\texttt{SIM}_{\mathrm{M4}}$ were assessed. The criterion for relevance judgments was made by comparing solution descriptions between a query incident in each query-set and each of k-top incident cases retrieved. The parameters used in $\texttt{SIM}_{\mathrm{M1}}$ - $\texttt{SIM}_{\mathrm{M4}}$ were the same as described in

Section 5.1. More specifically, the judgment is performed as the following: (1) Generate a random sample of 10 query incidents by each assessor; (2) Given every query incident in the sample, $\mathtt{SIM}_{\mathrm{M1}}$ - $\mathtt{SIM}_{\mathrm{M4}}$ are computed, and then for each of $\mathtt{SIM}_{\mathrm{M1}}$ - $\mathtt{SIM}_{\mathrm{M4}}$ the corresponding 10-top incident cases are retrieved; (3) For each of those retrieved incident cases, the assessor makes a assessment by selecting 'y' (relevant) or 'n' (irrelevant).

P is defined as the ratio of retrieved incident cases that are relevant, i.e., P = (the number of relevant items retrieved) / (the number of retrieved items (N_s)). R is defined as the ratio of relevant incident cases that are retrieved, i.e., R = (the number of relevant items retrieved) / (the number of relevant items (N_r)). In our experiment, $N_s = 10$ (10-top incident cases) and $N_r = |U|$, where U is the union set of the relevant incident cases among all of N_s incident cases returned by all $\mathtt{SIM}_{\mathrm{M1}}$ - $\mathtt{SIM}_{\mathrm{M4}}$. PR clearly trade off against one another: R is non-decreasing function of the number of incident cases retrieved, while P usually decreases as the number of incident cases retrieved is increases. A single measure that trades off P versus R is $\mathtt{F}_1$ measure, which combines into a single number as shown $\mathtt{F}_1 = \frac{2\mathtt{PR}}{\mathtt{P}+\mathtt{R}}$ [10]. In this respect, our effectiveness measure is only restricted to $\mathtt{F}_1$ measure. Every value measured by P, R, and $\mathtt{F}_1$ falls in the range [0,1], with 1.0 being the best score.

The result of the assessments measured by the 3 assessors are shown in Fig. 3. In every graph in the figure, M1 - M4 correspond to $\mathtt{SIM}_{\mathrm{M1}}$ - $\mathtt{SIM}_{\mathrm{M4}}$, respectively. In addition, the horizontal axis represents the 10 query incidents tested, and the vertical line denotes $\mathtt{F}_1$ values computed. Moreover, the 3 sets of averaged values of P, R and $\mathtt{F}_1$ of the 3 assessments are also shown in Table 4.

Table 4: The averaged assessment result.

	Assessor1			Assessor2			Assessor3			Total Average		
	P	R	$\mathtt{F}_1$	P	R	$\mathtt{F}_1$	P	R	$\mathtt{F}_1$	P	R	$\mathtt{F}_1$
$\mathtt{SIM}_{\mathrm{M1}}$	.740	.622	.670	.550	.499	.495	.710	.616	.652	.667	.579	.606
$\mathtt{SIM}_{\mathrm{M2}}$	.800	.671	.724	.630	.568	.568	.740	.642	.680	.723	.627	.657
$\mathtt{SIM}_{\mathrm{M3}}$	.770	.646	.696	.620	.556	.557	.770	.675	.711	.720	.626	.655
$\mathtt{SIM}_{\mathrm{M4}}$	.870	.731	.788	.700	.621	.629	.840	.735	.775	.803	.696	.731

To analyze the $\mathtt{F}_1$ values shown in Fig. 3, we performed *paired t-tests* at the 95% confidence level to assess whether the $\mathtt{F}_1$ values that are discovered by $\mathtt{SIM}_{\mathrm{M1}}$ - $\mathtt{SIM}_{\mathrm{M4}}$ are statistically different from each other. The paired t-test evaluates the significance of the difference between means of two independent data sets [18]. Referring to Fig. 3 and Table 4, we analyze the result of the assessments as follows:

(1) As seen in the figure, $\mathtt{SIM}_{\mathrm{M2}}$, $\mathtt{SIM}_{\mathrm{M3}}$ are proved to be more effective than $\mathtt{SIM}_{\mathrm{M1}}$, since their all $\mathtt{F}_1$ values are higher than those of $\mathtt{SIM}_{\mathrm{M1}}$, only except for the 3^{th}, 9^{th} query incidents in the 3^{th} graph (in the order from the top). In the 1^{th} graph, according to paired t-tests, $\mathtt{SIM}_{\mathrm{M2}}$'s $\mathtt{F}_1$ values were significantly higher than those of $\mathtt{SIM}_{\mathrm{M1}}$, while insignificant between $\mathtt{SIM}_{\mathrm{M1}}$, $\mathtt{SIM}_{\mathrm{M3}}$. In the 2^{th} graph, both $\mathtt{SIM}_{\mathrm{M2}}$, $\mathtt{SIM}_{\mathrm{M3}}$ were significantly higher than $\mathtt{SIM}_{\mathrm{M1}}$, while the both were insignificantly higher than $\mathtt{SIM}_{\mathrm{M1}}$ in the 3^{th} graph. But, on average, all the $\mathtt{F}_1$ values of $\mathtt{SIM}_{\mathrm{M2}}$, $\mathtt{SIM}_{\mathrm{M3}}$ are higher than those of $\mathtt{SIM}_{\mathrm{M1}}$ in all the assessments as seen in the table. This indicates that exploiting the additional meta knowledge, workgroup and incident classification, is useful in improving the effectiveness of similarity measurement;

(2) It is noted that $\mathtt{SIM}_{\mathrm{M2}}$, $\mathtt{SIM}_{\mathrm{M3}}$ may be similar to each other with regard to their effectiveness. Although, as seen in the table, on '`Total Average`', $\mathtt{SIM}_{\mathrm{M2}}$'s $\mathtt{F}_1$ value is slightly higher than that of $\mathtt{SIM}_{\mathrm{M3}}$, but $\mathtt{SIM}_{\mathrm{M3}}$'s $\mathtt{F}_1$ value is higher than that of $\mathtt{SIM}_{\mathrm{M2}}$ on 'Assessor3'. In fact, in paired t-tests on the $\mathtt{F}_1$ values between $\mathtt{SIM}_{\mathrm{M2}}$, $\mathtt{SIM}_{\mathrm{M3}}$ in all the 1^{th} - 3^{th} graphs, their differences were not statistically significant.

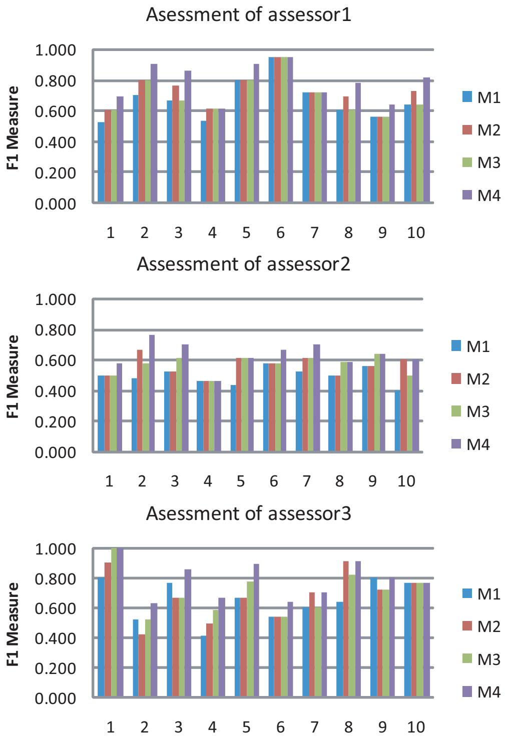

Figure 3: The result of $\mathtt{F}_1$ values from the 3 assessments.

(3) It is shown that $\mathtt{SIM}_{\mathrm{M4}}$ is the most effective among $\mathtt{SIM}_{\mathrm{M1}}$ - $\mathtt{SIM}_{\mathrm{M4}}$. In fact, according to paired t-tests, $\mathtt{SIM}_{\mathrm{M4}}$ was significantly higher than all of $\mathtt{SIM}_{\mathrm{M1}}$ - $\mathtt{SIM}_{\mathrm{M3}}$ in all the assessments with regard to $\mathtt{F}_1$ values. This result indicates a clear evidence that our similarity measure incorporating the combined information of workgroup, incident classification and incident description outperforms the similarity measures using just the partial combinations of them.

To help understand more about the experimental results, we present the following additional things examined based on Fig. 3 and Table 4: (1) As seen in the table, we can notice some differences of $\mathtt{F}_1$ values between the assessors. The reason is that the assessors may have different criteria when judging the relevance (or irrelevance) of incident cases retrieved for a given query incident. But, despite of the differences, our result shows that the relationship of "$\mathtt{SIM}_{\mathrm{M4}}$ > $\mathtt{SIM}_{\mathrm{M2}}$, $\mathtt{SIM}_{\mathrm{M3}}$ > $\mathtt{SIM}_{\mathrm{M1}}$" is consistently maintained in all the assessments, with regard to $\mathtt{F}_1$ values; (2) As seen the figure, there are also some differences between the $\mathtt{F}_1$ values in the same measures. For example, considering $\mathtt{SIM}_{\mathrm{M4}}$ in the 2^{th} graph, the $\mathtt{F}_1$ value of the 4^{th} query incident (q_4) is .462, while that of the 1^{th} query incident (q_1) in the 3^{th} graph is 1.0. Such a difference was occurred mainly due to the gap of the number of relevant incident cases n, existed in the CB. In the former case, $n = 3$ (i.e., P=.3), while $n = 10$ (i.e., P=1.0) in the latter case, but both the cases have the same R value measured as 1.0. Actually, we found that there were only 3 relevant incident cases for q_4 in the CB, while more than 10 for q_1. Thus, depending on the real number of relevant incident cases stored in the CB, the differences between the F_1 values in the same measures were examined.

6. CONCLUSION

This paper presented a new knowledge-rich similarity measure to improve the manual-based typical incident resolution process of incident management. Unlike similarity measures used in CBR, which only exploit incident (or problem) description, this measure incorporates additional two types of meta knowledge, workgroup and incident classification, to improve the capability of similarity measurement. Moreover, this measure exploits as much semantic knowledge as possible about various information contained in the incident cases. Based on this similarity measure, the most similar incident cases are retrieved for a new incident, and then the solutions contained in the retrieved incident cases can be used to help to generate the appropriate solution for the new incident. In the evaluation, we presented the effectiveness, technical coherence and feasibility of our similarity measure using a real dataset, based on the metrics of precision and recall. Our future plan is to evaluate this measure on fairly large and diverse datasets to improve its high-confidence. Currently, we are interested in modeling dynamic associative knowledge between non-compatible information individuals (e.g., between incident classification, workgroup, and incident description). By exploiting this knowledge, we envision that the performance of our similarity could be more improved.

7. REFERENCES

[1] HP Service Manager, http://www.hp.com/software.

[2] IT Infrastructure Library, "ITIL Service Delivery" and "ITIL Service Support". *OGC, UK*, 2003.

[3] O. Alm, E. Hyvönen, and A. Vehviläinen. OPAS : An ontology-based library help desk service. *4th European Semantic Web Conference 2007 (ESWC 2007)*, 2007.

[4] C. Bartolini, C. Stefanelli, and M. Tortonesi. Symian: A simulation tool for the optimization of the it incident management process. *Managing Large-Scale Service Deployment, LNCS*, pages 83–94, 2008.

[5] R. Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, November, 2002.

[6] K. H. Chang, P. Raman, W. H. Carlisle, and J. H. Cross. A self-improving helpdesk service system using case-based reasoning techniques. *Computers in Industry*, 30(2):113–125, 1996.

[7] P. Ganesan, H. Garcia-Molina, and J. Widom. Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems*, 21(1):64–93, 2003.

[8] C. Gliedman. Transitioning from incident to problem management: Key issues and challenges. *Forrester*, 2006.

[9] M. H. Göker and T. Roth-Berghofer. The development and utilization of the case-based help-desk support system homer. *Engineering Applications of Artificial Intelligence*, 12(6):665–680, 1999.

[10] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5 – 53, 2004.

[11] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *In Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, 1997.

[12] Y. F. D. Law, S. B. Foong, and S. E. J. Kwan. An integrated case-based reasoning approach for intelligent help desk fault management. *Expert Systems with Applications*, 13(4):265–274, 1997.

[13] R. Leopoldi. It service management - incident/problem management methods and service desk implementation best practices. *White Paper of RL Information Consulting LLC.*, 2003.

[14] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: An online lexical database. *International Journal of Lexicography (special issue)*, 3(4):235–312, 1990.

[15] T. Pedersen, S. V. S. Pakhomov, S. Patwardhan, and C. G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *J. of Biomedical Informatics*, 40(3):288–299, 2007.

[16] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Trans.Syst.Man Cybern*, 19(1):17–30, 1989.

[17] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research.*, 11:95–130, 1999.

[18] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *SIGIR '05*, pages 162–169, 2005.

[19] N. Seco, T. Veale, and J. Hayes. An intrinsic information content metric for semantic similarity in wordnet. *Techn. report, University College Dublin, Ireland*, 2004.

[20] A. Stahl. Learning of knowledge-intensive similarity measures in case-based reasoning. *PhD thesis, Technical University of Kaiserslautern*, 2003.

[21] A. S. Thomas Gabel. Exploiting background knowledge when learning similarity measures. *Proceedings of the 7th European Conference on Case-Based Reasoning (ECCBR 2004)*, 2004.

[22] C. J. van Rijsbergen. In *Information Retrieval. 2nd ed.*, 1979.

[23] I. Watson. Case-based reasoning is a methodology not a technology. *Knowledge-Based Systems*, 12(5-6):303 – 308, 1999.

[24] W. Wilke, M. Lenz, and S. Wess. Intelligent sales support with cbr. In *Case-Based Reasoning Technology, From Foundations to Applications*, pages 91–114, London, UK, 1998. Springer-Verlag.

[25] Q. Yang and J. Wu. Enhancing the effectiveness of interactive case-based reasoning with clustering and decision forests. *Applied Intelligence*, 14(1):49–64, 2001.