Knowledge-Based Systems 71 (2014) 211–226

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Effect of thesaurus size on schema matching quality ☆

Thabit Sabbah [a], Ali Selamat [b,*], Mahmood Ashraf [c], Tutut Herawan [d]

[a] Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
[b] UTM-IRDA-COE & Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
[c] Department of Computer Science, Federal Urdu University of Arts, Science and Technology, Islamabad, Pakistan
[d] Department of Information System, University of Malaya, 50603 Pantai Valley, Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 11 February 2014
Received in revised form 26 June 2014
Accepted 4 August 2014
Available online 17 August 2014

Keywords:
Schema matching
Thesaurus
Information Retrieval
Searching
Performance
Text similarity
Structured vocabulary

ABSTRACT

Thesaurus is used in many Information Retrieval (IR) applications such as data integration, data warehousing, semantic query processing and schema matching. Schema matching or mapping is one of the most important basic steps in data integration. It is the process of identifying the semantic correspondence or equivalent between two or more schemas. Considering the fact of the existence of many thesauri for identical knowledge domain, the quality and the change in the results of schema matching when using different thesauri in specific knowledge field are not predictable. In this research, we studied the effect of thesaurus size on schema matching quality by conducting many experiments using different thesauri. In addition, a new method in calculating the similarity between vectors extracted from thesaurus database is proposed. The method is based on the ratio of individual shared elements to the elements in the compound set of the vectors. Moreover, we explained in details the efficient algorithm used in searching thesaurus database. After describing the experiments, results that show enhancement in the average of the similarity is presented. The completeness, effectiveness, and their harmonic mean measures were calculated to quantify the quality of matching. Experiments on two different thesauri show positive results with average Precision of 35% and a less value in the average of Recall. The effect of thesaurus size on the quality of matching was statically insignificant; however, other factors affecting the output and the exact value of change are still in the focus of our future study.

1. Introduction

For more than two decades, thesauri were exploited in many IR applications. For example, it were used in web document classification [1], summarization [2], indexing [3], and in calculating the semantic similarity of documents written in the same or in different languages [4]. Thesaurus was also utilized to solve the problem of schema matching [5–7]. Recently, thesaurus is used to predict query difficulty in medical domain. It was concluded that the performance of the predictor is influencing with many factors such as the coverage of thesaurus or query mapping quality [8]. Earlier studies assumed that there are no general thesauri such that sufficient coverage are available, so that the use and impact of thesaurus was not studied widely [8]. However, a high quality thesaurus is available for some specific domains, also many thesauri with different coverage abilities and sizes are found in the same domain.

Such as any other controlled vocabularies, thesaurus is reusable and replaceable (i.e. can be reused in many different applications and can be replaced by another compatible thesaurus). However, the quality of the thesaurus is crucially to be assessed before reuse or replacement. According to [9] the size of the vocabulary is one of the main quality issues considered in measuring the quality of the controlled vocabulary. This research is discuss the effect of the thesaurus size on the quality of schema matching, thus, measuring and assessing of the thesaurus quality is out of this research's scope, details on thesaurus quality assessment can be found in [9,10].

Domain specific thesaurus are preferred to the common thesaurus such as WordNet in this research because of the common thesaurus are already used in this field as shown in the next

☆ This is an extended paper that has been presented in Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference by Sabbah, Thabit; Selamat, Ali, "Thesaurus Performance with Information Retrieval: Schema Matching as a Case Study," Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference, pp. 4494,4498, 13–16 October 2013.

* Corresponding author.
E-mail address: aselamat@utm.my (A. Selamat).

http://dx.doi.org/10.1016/j.knosys.2014.08.002
0950-7051/

paragraphs, moreover this research is studying the effect of the size of domain specific thesauri for single domain.

In information and database systems, schema is stands as the set of formulas (collection of meta-data) imposed on the data in the database. These formulas (also called integrity constraints) are applied to ensure the compatibility and describe the organization and the relations between database's parts and entities [11].

The importance of studying the effect of thesaurus size is coming from the vital need of effective and complete automatic solutions, because of the rapid expansion of application areas in which thesaurus and other vocabulary tools can be utilized such as natural language processing and Information Retrieval. For instance, schema matching forms the first and the crucial step toward data integration, however, the multiplicity of the obtainable common and domain specific vocabulary and linguistics tools that can be used, makes it hard to prefer one tool over others since the influences of tool's features such as size and coverage are not predetermined.

1.1. Schema matching related works

Schema matching, which is the process of identifying the semantic correspondence, or finding the equivalent elements between two or more schemas, is still an open research area since more than two decades. This is not only because schema matching is one of the basic operations [12] in many applications such as data integration, data warehousing, and semantic query processing, but also because it is an increasingly important problem itself [13], and as well as the uncertainty in the results of schema matching techniques [14,15]. Many approaches and tools were used to solve the problem of schema matching such as Cupid [16], LSD [17], and Corpus [18]. In addition, many surveys and classifications were published [19,20]. Few features of matching process were not in the focus of proposed approaches, and aspects such as structural, element, linguistics, and data model were discussed widely. Following is a summarization of the techniques used in schema matching approaches.

Many techniques were employed to carry out matching process; Machine-learning techniques were used in [17], learner-based approaches contains learner modules and specific module to direct learners. These approaches use neural networks advantages to find out the similarity between data sources. In [21] the object-oriented characteristics were exploited to determine the mapping between data sources' attributes. The problem of matching is not solved using this approach as well many proposed works using metadata; however, it is shifted into another problem, which is the problem of ontology mapping. Most of current schema matching tools use rules to carry out the matching, by using information such as elements names and descriptions, data types, hierarchy structure, and constraints. They are employed in determining the similarity at either element level or schema level [16,21,22].

Most effective rule-based schema matching methods usually consist of three phases; linguistic, constraint-based, and structural matching [23]. In linguistic phase, methods depend on string matching in general to find out the similarity between elements names. Current schema matchers usually use WordNet, a large lexical database of English [24] to consider the semantic relationships between elements labels [6]. However, it is common that algorithms in this category use combined methods to get high computed similarity, methods of label normalization to improve schema matching was also by [6,7]. Cupid matcher exploits linguistic matching in a comprehensively and efficiently manner to produce high similarity [16]. Incorrect results that are obtained from linguistic matching phase are usually adjusted in constraint-based matching phase. Data type constraint, data types' compatibility measurement method are usually used as the initial solution of incorrect or ambiguous results of linguistic matching phase [25,16]. Structural matching phase is used to solve the problems of context similarity, these problems are generally appear in XML schema matching where the structure document and the constraints on nodes and edges differs from rational schemas [23] describes such problems in details.

Based on the conclusion of [8], this paper studies the effect of thesaurus size (in aspects of number of terms, number of lead-in terms, and number of cross relations) on the results of schema matching using thesaurus.

1.2. Research contributions

Although there are few exiting works in the thesaurus based schema matching field, the main contributions of this research encompass:

- Presenting an experimental study of the effect of thesaurus size on schema matching quality. Three agricultural thesaurus of different size are utilized and compared, and the results are evaluated through several objective functions.
- A new method to compute the similarity between vectors extracted from the thesaurus is proposed.
- Moreover, this paper explains in detail many of the technical aspects to be considered when using thesaurus.
- The experimental results shows that the effect of thesaurus size in the quality of matching is statistically insignificant. However, an increment in the average of similarity with distinctive values are recorded.

1.3. Research limitations

This research is studying the effect of thesaurus size on the quality of schema matching, by utilizing three thesauri from the Agriculture domain to carry out the matching process on the element level, and the results are analyzed in many different perspectives. Therefore, some other perceptions such as thesaurus construction and evaluation, results (Precision, Recall, and F-measure) optimization, and the method complexity are not in the scope of this research.

In the rest of this paper, Section 2 explains the methodology. Section 3 presents the study setup. Section 4 shows the results as well as a discussion of these results. Finally, this work is concluded in Section 5.

2. Schema matching based on linguistic analysis with thesaurus

This paper studied the impact of thesaurus size on the quality of schema matching. The applied methodology is based on exploiting thesaurus to carry out the matching process. Fig. 1 shows the methodology framework, and the next subsection explains it in details.

The method consists of three main phases as shown in Fig. 1. Numbers is circles 1, 2 and 3 represent these phases. In phase one, two schemas (S_x and S_y) are part of the input of the (Apply Thesaurus) process, thesaurus is the other part of input for this process, and the output of (Apply Thesaurus) process are two sets of vectors of terms (S_x mass and S_y mass). These two sets of vectors will form the input of phase two, which is (Calculating Similarity Matrix) to produce the Similarity Matrix (SM) between the schemas' elements; The third phase is (Extracting the Final Mapping) that uses SM as an input to generate the final mapping list.

Algorithms and details of these phases are explained in following sub-sections.

2.1. Methodology

As shown in Fig. 1, thesaurus is utilized in solving the problem of schema matching at the element level based on textual analysis of elements' descriptions (definitions) of input schemas (Schema One and Schema Two). Each input schema contains number of elements, for abbreviation and algorithms writing purposes these schemas are referred as S_x where $x \in \{1,2\}$. Moreover, the number of elements in these schemas is referred as n and m. Following is a detailed description of the three phases of the method.

2.2. Phase one

This phase includes many pre-processing steps such as removing stop words, removing numbers, and characters not matching with thesaurus language and content. The main process in this phase is (Applying Thesaurus). The output of this phase is two sets of vectors of terms (masses) where each vector represents one element in the schemas.

Apply thesaurus process: in this process, thesaurus is applied on elements' textual descriptions, one by one for both schemas S_1 and S_2. Applying thesaurus means searching for every word from the text (i.e. element description) into thesaurus database and retrieving the related terms from thesaurus, to build up the mass of terms related to the word being processed; this mass is denoted by $mass_w$ in the Algorithm 1.

Algorithm 1. Applying thesaurus on element description algorithm

Input: $S_1 = \{(e, desc)_{10},\ldots,(e, desc)_{1n}\}$ // e: element name | desc: element description
$S_2 = \{(e, desc)_{20},\ldots,(e, desc)_{2m}\}$
For ($S_j \in \{S_x, S_y\}$) loop // loop through the schemas
$S_j mass \leftarrow \{\}$ // initialize set of schema element_masses' set
For ($e_k \in S_j$) loop // loop through the elements in the schema, k= 0 .. n|m
$element\_mass_k \leftarrow \{\}$ // initialize element_mass (vector)
For (word $\in$ element s_{jk} description) loop// loop through the words in the description
If (word found in thesaurus Index)
$mass_w \leftarrow$ ***get_related_terms***(w) // retrieve all terms from thesaurus database related to term (w)
$element\_mass_k \leftarrow \cup\ mass_w$
End If //
End loop// through the words in the description
End loop // through the elements in the schema
$S_j mass_k \leftarrow (e_k, element\_mass_k)$
End loop // through the schemas
Output: S_1 mass $=\{(e, element\_mass)_{10},\ldots,(e, element\_mass)_{1n}\}$
$S_2 mass = \{(e, element\_mass)_{20},\ldots,(e, element\_mass)_{2m}\}$

Different masses $mass_w$(s) are then accumulated on the element level into one mass (*element_mass*) that represents the Result of Applying Thesaurus (RAT) on the element e_i of the schema ($RATe_iS_x$) as shows in Algorithm 1. This phase contains extensive searching processes because the process of Applying Thesaurus is done for every term in every description in both schemas, term may be one word or multiple word that is known also as Compound Term. The searching algorithm applied in this phase is explained in Section 2.5, and the function (get_related_terms(w)) which used to retrieve all terms related to term (w) from the database is explained in Section 2.6.

2.3. Phase two

In this phase, the two vectors resulted from previous phase are used as the input of (Calculating Similarity Matrix) process. Similarity between Result of Applying Thesaurus (RAT) of each element from S_1 with all RATs of elements of S_2 were calculated to generate the similarity matrix; Algorithm used in calculating similarity matrix is shown in Algorithm 2.

Algorithm 2. Calculating similarity matrix algorithm

Input: $S_1 mass = \{(e, RATe_1S_1)_0,\ldots,(e, RATe_nS_1)_n\}$
$S_2 mass = \{(e, RATe_1S_2)_0,\ldots,(e, RATe_mS_2)_m\}$
SimMatrix←Matrix[n][m]
Initialize SimMatrix; // set all cells to 0
for ($e_i \in S_x mass$)// i = 0 .. n
for ($e_j \in Smass_y$) // j = 0 .. m
$SimMatrix_{ij} \leftarrow$ ***Similarity***($RATe_iS_x$, $RATe_jS_y$)
Output: SimilarityMatrix[n][m]

The ***Similarity*** between two elements is defined based on the following equation:

$$Similarity\ (e_iS_x, e_jS_y) = \left(\frac{RAT(e_iS_x) \cap RAT(e_jS_y)}{RAT(e_iS_x) \cup RAT(e_jS_y)}\right), \quad (1)$$

where *RAT* is Result of Applying Thesaurus on element. The similarity in Eq. (1) considers the vectors as sets of elements where duplicate elements is not allowed. Since the vectors represents all terms from thesaurus related to the element ($e_i|e_j$) of schema ($S_x|S_y$), then the frequency of terms is not considered since one term from thesaurus may appears in the results vector because it is related with many others terms with different relationships. Moreover, the interest of the similarity measure in Eq. (1) is the differences between the two masses of terms extracted from the thesaurus for certain text. Unlike some other similarity measurements such as cosine similarity where the frequency of terms is considered or the frequency of errors (mismatched elements) such as in hamming distance measurement, in our proposed similarity equation the existence or absence of the terms in the mass is the main concern of this measure because of the above mentioned reason.

Fig. 2 shows an example of calculating similarity between two elements.

Similarity is calculated between all possible elements pair's combinations, and stored in the Similarity Matrix.

For evaluation purposes, the ***Similarity*** between two element's descriptions is also calculated using the common cosine similarity equation [26]. The cosine similarity between two vectors (e_iS_x, e_jS_y) is defined as follows:

$$\text{cosine similarity}\ (e_iS_x, e_jS_y) = \frac{e_iS_x \cdot e_jS_y}{\|e_iS_x\|\|e_jS_y\|} = \left(\frac{\sum_{w=1}^{n} e_{wi}S_x \times e_{wj}S_y}{\sqrt{\sum_{w=1}^{n} e_{wi}S_x} \times \sqrt{\sum_{w=1}^{n} e_{wj}S_y}}\right), \quad (2)$$

where e_iS_x, e_jS_y are the vectors resulting from Applying Thesaurus on element i of schema S_x, and element j of schema S_y respectively, and w is word in vector e.

The values in similarity matrix were normalized based on the following linear transformation formula:

$$X_n = \frac{X_0 - X_{\min}}{X_{\max} - X_{\min}},$$

where X_n = new X value (after normalization), X_0 = current value of X (before normalization), $X_{\min}$ = minimum value of X in the similarity matrix, and $X_{\max}$ = maximum value of X in the similarity matrix.

2.4. Phase three

In this phase, the Similarity Matrix (SM) generated from phase two was used as an input for (Extract Final Mapping) process that generates the final mapping set. The maximum and second maximum value approach [27] was applied in extracting the final mapping as shown in Algorithm 3.

Algorithm 3. Calculating similarity matrix

```
1:  Input:
2:    S = SimilarityMatrix[n][m]
3:  Variables:
4:    cellIndex=[row,col]
5:    finalMapping={}// set of cell Indexes
6:  While S contains value > 0
7:    max ← getMaxValue(S) //get the maximum value in the matrix
8:    cellIndex(row,col) ← getRowCol(max) // get the x,y index of max value in the matrix
9:    If (max is unique) // check for uniqueness of the Max value
10:     FinalMapping ← U cellIndex(row,col) // append cell index to the final Mapping list
11:     S[row] ← 0 // set similarity value to zeros in the row
12:     S[col] ← 0 // set similarity value to zeros in the column
13:     S[row,col] ← −1 ∗ max // set max value to negative in the similarity matrix
14:   Else
15:     ∀ (S[row,col] = max): S[row,max]←0 // set all cells equals to max to zero
16:   End If
17: Loop // while
18: Output:
19:   finalMapping
```

In this algorithm, a matching (mapping) between two elements (one in the header of the row and other in the header of the column) is considered if the similarity value in the cross cell is the maximum value in the matrix. Then all values in the row and the column were set to zero. This process will be repeated until all similarity values in the matrix become zeroes or less than the threshold value. The problem of this criterion will come up when the maximum value is not unique in the similarity matrix and more than one of maximum value occurrences found in the same row or the same column, this case requires us to check the second maximum value of the matrix where the second maximum value is considered as the mapping.

2.5. Searching thesaurus database

Searching thesaurus database is one of the main processes performed in all applications that use thesauri either in the core or as an auxiliary tool. In this research thesaurus is used as the core of the matching process. Thesaurus was applied on all elements' textual descriptions. The procedure ***get_related_terms*** (mentioned in Algorithm 1) contains extensive searching processes in thesaurus database, because the need to search for every term from text into the thesaurus database. The term may be one word or multiple words (also called compound term), although the thesaurus contains one word terms and compound terms too. The direct approach to deal with such case is the brute force method in which the text is traversed by considering the term as one word in first round, and then the traversing is repeated by considering the term as double word, and so on. Traversing of the text will stop when the number of words in the term from the text exceeds the number of the words in the longest term in the thesaurus database. This brute force algorithm is the less efficient search algorithm [28]. An efficient searching algorithm [29] is applied to carry out this process. Algorithm 4 shows the applied algorithm used to reduce time required for searching text into thesaurus database. This algorithm is discussed in details [29].

Algorithm 4. Searching text into thesaurus database

```
1:  for (w ∈ text)
2:    if (w found in Index)
3:      termsLengths ← getTermsLengthsThatStartsWith(w)
4:      for (l ∈ termsLengths)
5:        compoundTerm = buildCompoundTermfromtextoflength(l)
6:        if (compoundTerm found in DB)
7:          addRelatedTermsToResultSet
8:        endif compoundTerm found in DB
9:      end for length
10:   end if w found in Index
11: end for w
```

The main idea of Algorithm 4 is to search for the word (w) into the index vector of the thesaurus instead of search for the word (w) into the terms' table of the thesaurus that surely contains many compound terms. Index vector of the thesaurus is a vector that contains the distinctive first token of terms or compound terms of the thesaurus. Two benefits are gained from this step: **First**, once (w) is found in the index, then for sure there is one or more raw (one word term or compound term) in thesaurus starts with that word. Otherwise, there is no need to look into thesaurus for any compound term that starts with the word (w); because for sure there is no compound term starts with that specific word. **Second**, as a result of finding (w) in the index, the set of lengths of the compound terms in thesaurus that starts with (w) – *step number three in algorithm 3* – can be defined, so that the list of compound terms of the required lengths from the text starting from the word under consideration could be built up.

2.6. Retrieving term mass from thesaurus database

Finally, once the term is found in the thesaurus database, as mentioned in Algorithm 1, the function ***get_related_terms***(w) is called to retrieve the *term mass* from the thesaurus database by

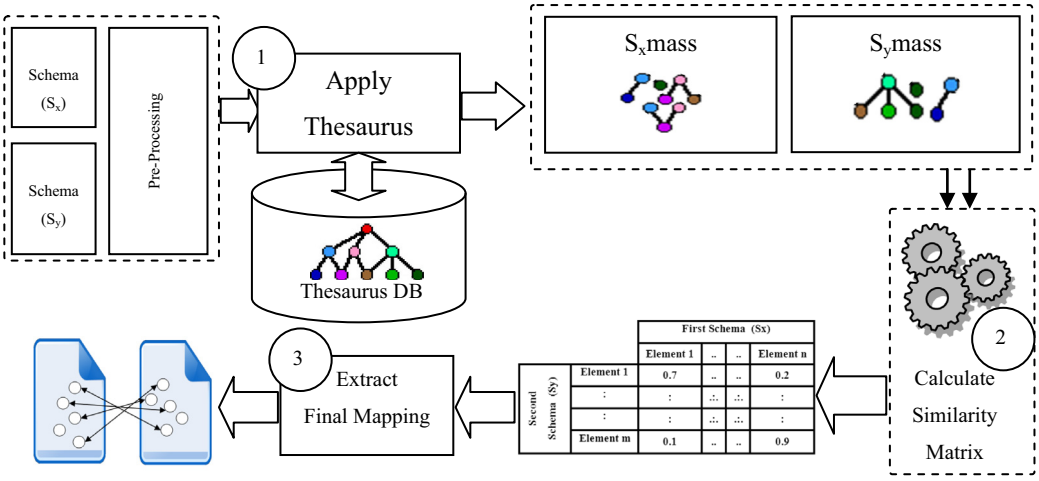

Fig. 1. Methodology framework.

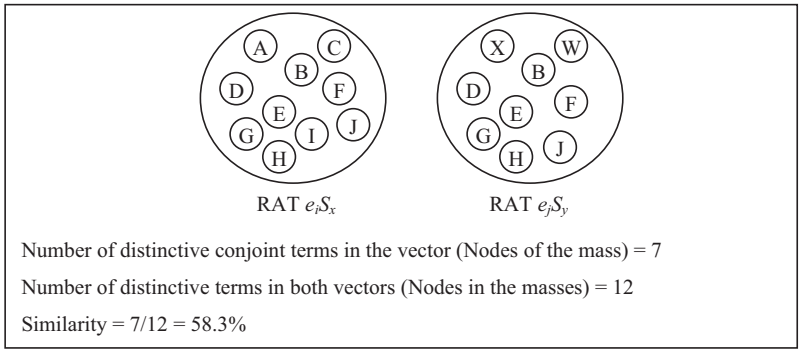

Fig. 2. Example of calculating similarity between two elements.

executing many hierarchical dynamic queries such the queries below. The term mass of a term is defined as all the terms in the database connected to the term with any of the thesaurus relations, which are Boarder terms, Narrow Terms, Related Terms, and the Preferred Terms.

```
SELECT a.r_term_code, b.term, b.term_tokens_count,
  b.is_nonprefered_term
FROM use_terms_relations a, terms b
WHERE a.r_term_code = b.term_code and          ... Query
  a.term_code = ?;                            (1)

SELECT a.term_code,x.term, a.r_term_code
  RT_Term_Code, y.term RT_Term
FROM rt_terms_relations a, terms x, terms y
WHERE a.term_code = x.term_code and            ... Query
  a.r_term_code = y.term_code and             (2)
  a.term_code = ?''.

SELECT a.term_code,x.term, a.r_term_code
  BT_Term_Code, y.term BT_Term, level
FROM bt_terms_relations a, terms x, terms y
WHERE a.term_code = x.term_code and a.r_term_code =
  y.term_code
START WITH a.term_code = ?
CONNECT BY PRIOR a.r_term_code =               ... Query
  a.term_code'';                              (3).

SELECT a.term_code,x.term, a.r_term_code
  NT_Term_Code, y.term NT_Term, level
FROM nt_terms_relations a, terms x, terms y
WHERE a.term_code = x.term_code and a.r_term_code =
  y.term_code
START WITH a.term_code = ?
CONNECT BY PRIOR a.r_term_code =               ... Query
  a.term_code'';                              (4).
```

The queries (1) and (2) are used to retrieve the PREFERED and related terms respectively by using the ordinary **SELECT** statement structure, however the queries (3) and (4) are hierarchal[1] (recursive) queries that retrieve the terms connected by the Boarder and Narrow relation.

2.7. Evaluation and ranking

The quality measures precision, recall, and F-measure as defined in [30] are used to evaluate the quality of schema matching with different thesauri. Precision, recall, and F-measure are used in IR domain, however it is commonly used to schema matching evaluation [6]. In addition, in the case of common matches between manual and automatic, the quality of overall similarity is compared based on two approaches; first, the comparison based on Maximum value, and second is the comparison based on the Average value to show the enhancement in the overall similarity of common matches among thesauri used.

To calculate precision, recall, and F-measure the manual matches generated be the domain expert as in [31] were considered, then for each experiment the set of true positives (TP), false

[1] http://docs.oracle.com/cd/B19306_01/server.102/b14200/queries003.htm

```
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema" targetNamespace="urn:books" xmlns:bks="urn:books">
  <xsd:element name="AGR102" type="xsd:string">
    <xsd:annotation>
      <xsd:documentation>Introduction to the basic concept and practices of horticulture.Emphasis is on the establishment, management and
          use of horticultureplants in the garden and home.Students will have hands-on experience while learning about seedlings, cuttings,
          potting and planting.
      </xsd:documentation>
    </xsd:annotation>
  </xsd:element>
  <xsd:element name="AGR115" type="xsd:string">
    <xsd:annotation>
      <xsd:documentation>This course is designed for students who understand the fundamentals of horse care and feeding and have some
          proficiency in Western riding. Major topics for the course include horse anatomy and conformation, health care, training, and
          advanced riding techniques.
      </xsd:documentation>
    </xsd:annotation>
  </xsd:element>
```

Fig. 3. Example of schema used in the experiments.

positives (FP), and false negatives (FN) were determined. Based on these sets the quality measures were calculated as follows:

$$\text{Precision} = \frac{|\text{TP}|}{|\text{TP}| + |\text{FP}|},$$

$$\text{Recall} = \frac{|\text{TP}|}{|\text{FN}| + |\text{TP}|}, \text{ and}$$

and

$$\text{F-measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}.$$

3. Study set-up

3.1. Domain

Many of previous studies on schema matching such as [16,32,33] use schemas from the domain of E-commerce. However there were many obstacles to use these schemas in this research; for example, these schemas do not include a textual description of its elements, and there are no thesauri available for E-commerce domain. So, data from agricultural domain were utilized as the dataset.

Agricultural knowledge domain has tremendously progressed for the past several decades.[2] Less information on the exact size of this knowledge domain is found. However, the agricultural information are represented in many machine-readable formats by different global organizations. The National Agricultural Library Thesaurus[3] (NALT) is a thesaurus developed by the National Agricultural Library (NAL) of the United States Department of Agriculture. When it released for the first time it contains 42,326 descriptors and 25,985 non-descriptors organized into 17 subject categories. Currently it contains more than 98,000 term and available in two languages (English and Spanish). AGROVOC is a multilingual thesaurus designed in early 1980s by Food and Agriculture Organization of the United Nations (AGROVOC Thesaurus[4]) to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains. The latest edition of AGROVOC contains over 32,000 concepts. The Chinese Agricultural Thesaurus[5] (CAT) is the largest agricultural thesaurus in China that maintained by AII of CAAS. It contains more than 63,000 concepts most of them have English translation.

3.2. Dataset

The dataset used in these experiments consists of two schemas. Each schema represents a set of 23 courses offered by a university. The courses data is represented as XML schema files (.xsd). Fig. 3 shows a part of the schema file.

In the schema file, each (<xsd:element) node represents one element with the name mentioned in the (name) property, and the node (<xsd:documentation>) contains the textual description of the element. The two sets of courses were tested to find the equivalent courses between them. For experimental uses, sets were named as follows:

Set one: $S_x = (e_{x0}, e_{x1}, e_{x2}, \ldots, e_{x22})$, and
Set two: $S_y = (e_{y0}, e_{y1}, e_{y2}, \ldots, e_{y22})$,

Courses' descriptions in both sets were processed and analyzed using different thesauri in the same domain, subsequent section explains more about the used thesauri.

3.3. Thesauri

Three agricultural thesauri were used. Two of them are different versions of the same thesaurus. These thesauri are The National Agricultural Library Thesaurus 2008 Edition (referred as NAL2008), The National Agricultural Library Thesaurus 2012 Edition (referred as NAL2012), and the thesaurus presented by Food and Agriculture Organization of the United Nations (referred as AGROVOC). All thesauri were downloaded from the Internet, and processed by special tools to meet experiment's environment.

3.3.1. Thesaurus pre-processing

NAL thesaurus as well as AGROVOC thesaurus are downloadable from their official websites in many different formats such as XML, RDF-SKOS, PDF, MARC, plain text for NAL Thesaurus and XML, SKOS, MYSQL, Protege DB, OWL and ISO2709 for AGROVOC thesaurus. The pre-processing of the thesauri depends on the used format accordingly, In This research the XML-SKOS format is used, a sample of thesaurus concept "Chamidae" is shown in Fig. 4 as it appear in the downloaded thesaurus of format XML-SKOS.

The thesaurus is transformed into rational database based on the British standards 8723 data model [34] and the extension of the model in [29]. Fig. 5 shows a part of the class diagram of thesaurus data model as in [34]:

[2] http://www.kfh.ch
[3] http://agclass.nal.usda.gov/
[4] http://aims.fao.org/standards/agrovoc/about
[5] http://cat.aii.caas.cn/

```
<skos:Concept rdf:about="http://aims.fao.org/aos/agrovoc/c_47856">
        <skos:inScheme rdf:resource="http://aims.fao.org/aos/agrovoc"/>
        <skos:broader>
                <rdf:Description rdf:about="http://aims.fao.org/aos/agrovoc/c_942">
                        <skos:narrower rdf:resource="http://aims.fao.org/aos/agrovoc/c_47856"/>
                </rdf:Description>
        </skos:broader>
</skos:Concept>
```

(a)

```
<CONCEPT>
        <DESCRIPTOR>Chamidae</DESCRIPTOR>
                <BT>Veneroida</BT>
                <NT>Arcinella</NT>
                <NT>Chama</NT>
                <ES>Chamidae</ES>
                <TNR>192793</TNR>
</CONCEPT>
```

(b)

Fig. 4. Thesaurus sample in XML-SKOS format.

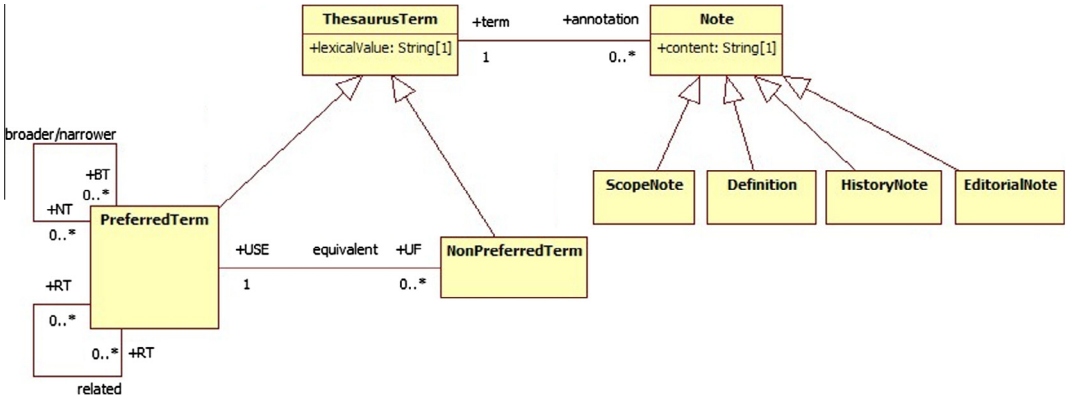

Fig. 5. Class diagram of thesaurus data model adapted from British Standards.

The general steps of pre-processing are summarized in the following steps:

Step 1: Extract and save terms identifiers.
Step 2: Extract and Save terms relations.
Step 3: Interconnect terms with extracted relations.
Step 4: Create terms index (for applying the efficient search method as in [29]).

3.3.2. Thesaurus statistics

Table 1 shows the main specifications of the thesauri used in the experiments of this research.

From Table 1, it seen that NAL2012 contains the largest number of terms, lead-in terms, and cross-relations, while AGROVOC has the least number of all specifications.

Other thesaurus specifications, such as the "Number of Words in Term" should also be considered in thesaurus performance measurement; this property influences the speed of calculating similarity. Fig. 6 shows the percentage of terms that contains one, two, three, and four or more words of terms in each thesaurus used in the experiments.

It can be seen from Fig. 6 that for all thesauri used, one word terms are less than 35% while the remaining terms are compound terms (i.e. terms consists of two or more words). As mentioned before the number of words in the term influences the speed of similarity calculating which means that an efficient algorithm is needed to carry out this job.

Table 1
Thesauri specifications.

	NAL2012 thesaurus	NAL2008 thesaurus	AGROVOC thesaurus
Total terms	87,438	69,794	40,623
Lead-in terms	38,418	30,212	22,508
Cross-relations	201,773	162,202	154,825

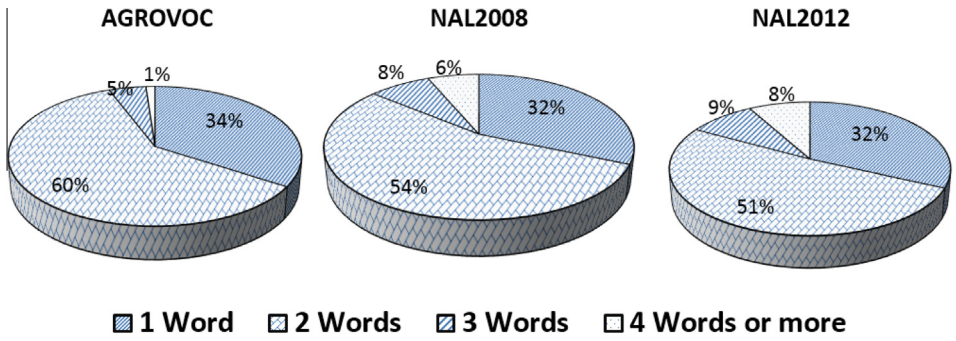

Fig. 6. Percentage of terms of one word terms and compound terms.

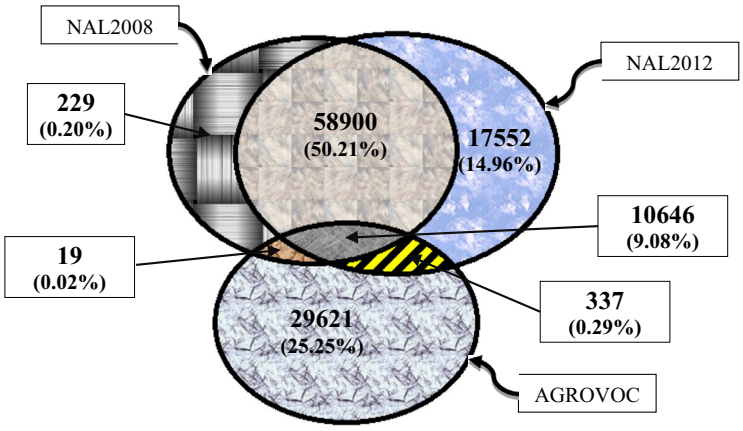

Fig. 7. Thesauri overlapping.

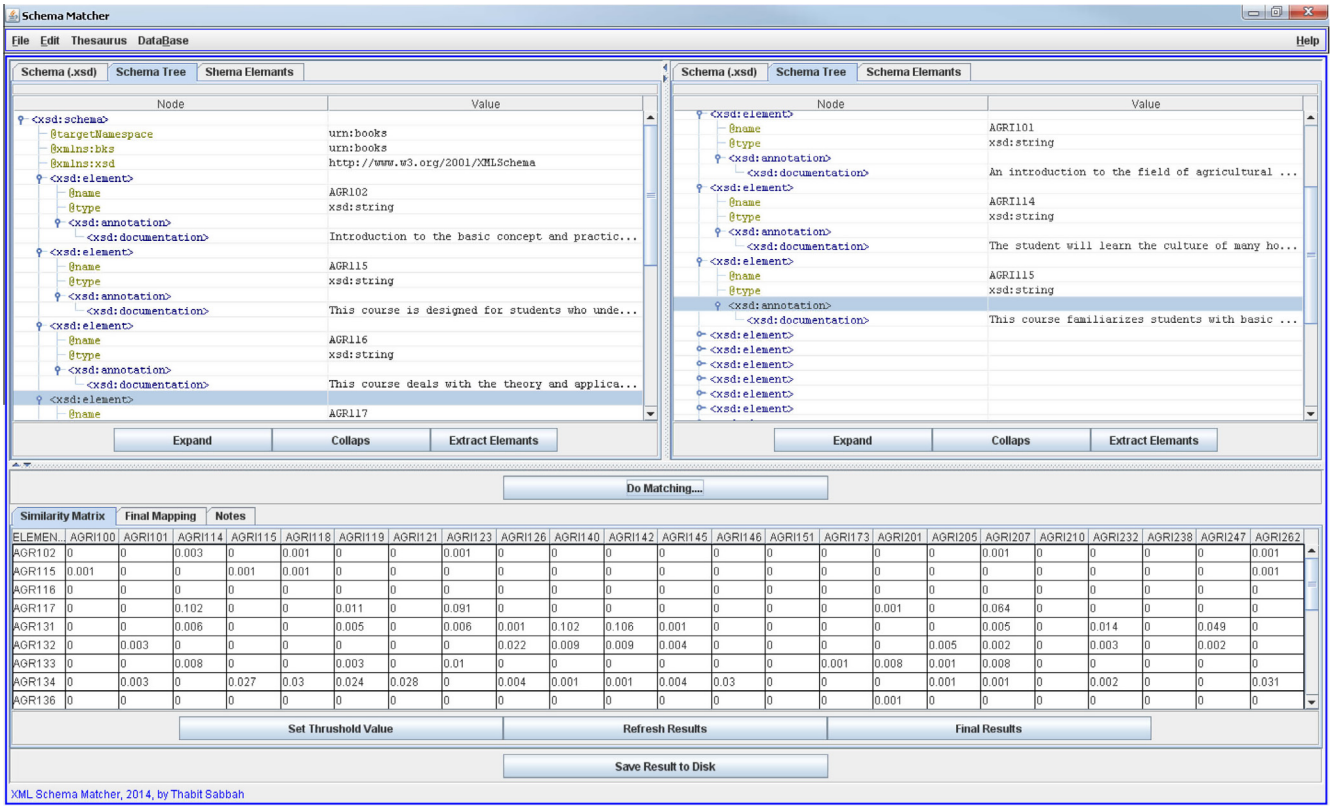

Fig. 8. GUI of the schema matching application.

Whereas different versions of the NAL thesaurus and the AGROVOC thesaurus are used in this study, these thesauri overlap with each other. Fig. 7 shows the number of overlapped terms and the ratios relative to the total number of distinctive terms in all thesauri.

The total number of distinctive terms in all thesauri is 117,304 terms. As shown in Fig. 7, the largest ratio of overlapping occurs between NAL2008 and NAL2012, which are different versions of the same thesaurus. However, the conjoint terms between all thesauri is near to 10% of the total number of terms. This study consider less attention to the influence of overlapping.

3.4. Experiment environment and application

To carry out the experiments, Oracle database with Java application developed especially for that purpose were installed. Fig. 8 shows the interface of the Java application.

The application has the facility to validate the loaded schemas, and to extract elements' names and their textual descriptions in a tree format before starting the matching process. The similarity matrix and the final mapping can be also saved to the file system.

4. Results

The two sets of courses used in the experiments were manually matched by an expert [31], results of manual and Automatic Matching of the experiments are shown in Table 2.

In Table 2 the similarity values in are based on Eq. (1) discussed in Section 2.3. The sub-table (a) represents the manual matches by domain Expert, and sub-table (b) represents the automatic matches based on NAL2008 thesaurus, while sub-tables (c) and (d) represent the automatic matches based on NAL2012 and AGROVOC thesaurus respectively. The matching results can be visualized as in Fig. 9.

In Fig. 9, the numbers on x-axis and y-axis represent the number of elements in schemas, while the bubbles represent the matches between elements, for example, there is a matching between element 5 from schema 1 and element 16 from schema 2 in manual matching. This matching is referred as pair (5,16) where the Pair stands for the two matched elements, and the numbers between brackets represents the number of elements in schema 1 and schema 2 respectively; the size of the bubble represents the value of similarity between the two elements. For matches that are common among manual matching and automatic ones, the bubbles appears to be over-lapping as for pairs (6,15) and (1,0) and others. The contingency table of the automatic results in relative to the manual matches are shown in Table 3.

Table 3 shows the number of matches' pairs distribution generated by each thesaurus relative to the manual matching. For example, in the experiment based on NAL2008, four pairs of matching elements are matched correctly by the automatic matcher, while 16 pairs are matched automatically incorrectly, and 6 pairs are incorrectly not matched. However, the number of pair in the cell of intersection of row total and column total represents the possible number of permutation of matches between schemas elements.

4.1. Discussion and analysis

This subsection discusses the results from many point views.

4.1.1. Discussion of precision, recall, and F-measures results

Precision, recall, and F measure for each experiment were calculated relative to manual matches, using the contingency table (Table 3) where the TP, FP, and FN sets are as follows:

Table 2
Matching similarities based on different thesaurus.

Element # in Schema 1	Element # in Schema 2	Similarity %
(a)		
11	18	25.00
14	1	81.00
15	13	100.00
16	8	100.00
18	9	100.00
3	2	63.00
4	7	38.00
5	16	94.00
6	15	50.00
7	11	0.10
(b)		
14	1	100.00
1	0	93.90
9	11	93.20
18	16	78.10
13	13	76.60
3	2	73.80
22	12	53.90
16	8	52.60
7	6	40.50
21	14	39.80
4	19	27.80
11	4	18.10
5	9	18.10
6	15	15.70
15	10	6.40
20	22	4.50
0	7	2.00
19	17	0.90
10	18	0.90
8	20	0.90
(c)		
14	1	100.00
1	0	94.90
22	11	93.00
12	16	82.40
16	8	78.80
0	2	77.50
13	13	74.60
3	7	48.50
7	12	39.90
18	14	32.30
10	18	28.10
5	9	22.60
4	10	19.30
6	15	7.30
20	22	4.90
9	19	1.70
21	17	1.50
19	6	0.90
11	4	0.90
8	20	0.90
15	5	0.60
(d)		
22	8	100.00
3	2	96.20
16	11	82.90
1	0	73.40
12	17	68.60
4	5	60.70
15	16	47.90
6	18	37.40
17	4	33.40
21	9	14.60
14	12	12.90
10	14	11.80
2	20	8.80
11	1	8.10
0	10	5.50
8	15	2.50
7	3	2.30
19	6	1.70
5	13	0.70
18	19	0.50

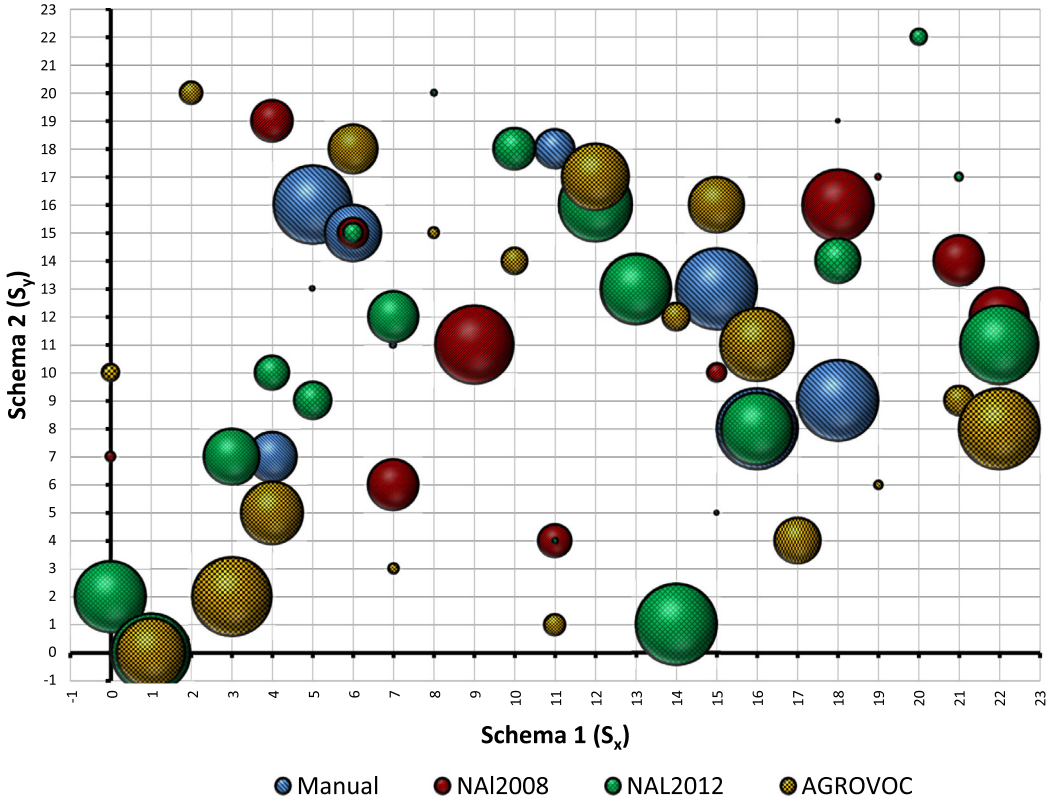

Fig. 9. Results of manual and automatic matching.

Table 3
Contingency table of automatic matching results relative to the manual results.

Automatic		Manual		Row total
		Matches	Non-matches	
NAL2012	Matches	3	18	21
	Non-matches	7	231	238
Column total		10	249	259
NAL2008	Matches	4	16	20
	non-matches	6	233	239
Column total		10	249	259
AGROVOC	Matches	1	20	21
	Non-matches	9	229	238
Column total		10	249	259

TP: the set of pairs matched manually and automatically.
FP: the set of pairs matched manually but not automatically.
FN: the set of pairs matched automatically but not manually.

Table 4 summarizes the results of Precision, recall, and F-measure for the experiments:

Two main remarks can be noticed from Table 4. One is the low precision, recall, and F-measure values. The proposed technique depends on searching for the words from elements' descriptions in the thesaurus. In the experiments the exact words are searched and no text pre-processing were applied, so the abbreviations, misspelled words, numbers written as words, inappropriate punctuations contained by the text will not contribute to the outcome of searching. For example, line 4 in Fig. 3 contains the expression (horticulture.Emphasis) which is considered as one word (because of no space between words), however, it will be recognized as two search terms if punctuations replacement is applied. To overcome this issue, some techniques can be applied such as text pre-processing, dictionary validation, punctuations replacement, and text expansion based on vocabulary tools.

Second, it can be seen that the use of rich thesaurus (in features), which is NAL2012, does not lead to higher precision and recall results. However, the use of AGROVOC thesaurus that has fewer terms, lead-in terms, and cross-relations cause a law precision and recall values. Fig. 10 shows the precision, recall, and F measure and the number of terms in each thesaurus.

Table 4
Precision, recall, and f measure for automatic matching.

	Nal2012	Nal2008	AGROVOC
Precision	0.30	0.40	0.10
Recall	0.15	0.20	0.05
F-measure	0.20	0.27	0.07

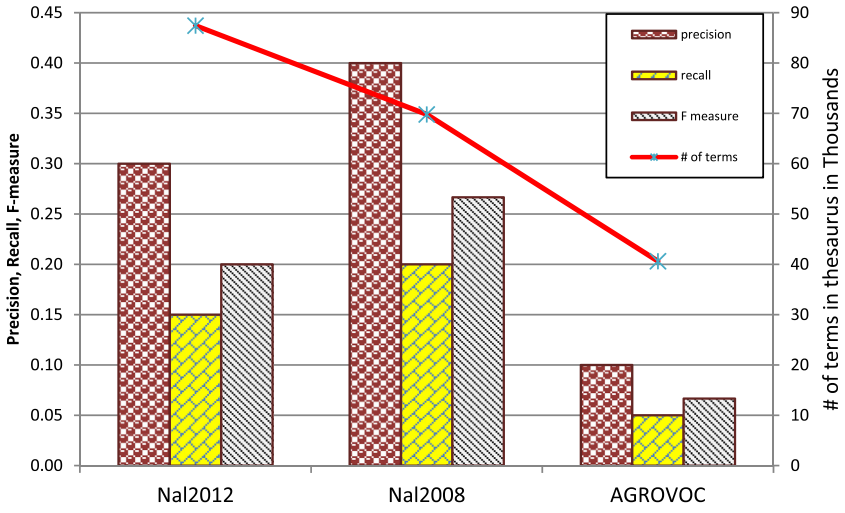

Fig. 10. Precision, recall, and F-measure measures for different thesauri.

Table 5
Common matches from results of using NAL2008 thesaurus and NAL2012 thesaurus.

Group no	Pair	Similarity		Absolute value of difference	Similarity avg.	Maximum similarity
		NAL2008 thesaurus	NAL2012 thesaurus			
1	(1,0)	0.939	0.949	0.010	0.944	0.949
2	(10,18)	0.009	0.281	0.272	0.145	0.281
3	(11,4)	0.181	0.009	0.172	0.095	0.181
4	(13,13)	0.766	0.746	0.020	0.756	0.766
5	(14,1)	1.000	1.000	0.000	1.000	1.000
6	(16,8)	0.526	0.788	0.262	0.657	0.788
7	(20,22)	0.045	0.049	0.004	0.047	0.049
8	(5,9)	0.181	0.226	0.045	0.204	0.226
9	(6,15)	0.157	0.073	0.084	0.115	0.157
10	(8,20)	0.009	0.009	0.000	0.009	0.009
Average		0.381	0.413	0.087	0.397	0.441
Enhancement in Similarity of MAX approach relative to Avg.		0.059	0.028		0.043	

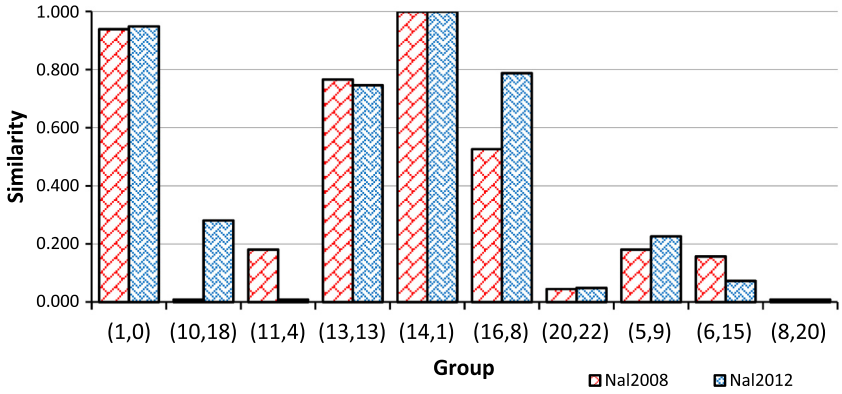

Fig. 11. Similarity values of common matches between NAL2008 and NAL2012.

As seen from Fig. 10, the precision was the least in case of using AGROVOC thesaurus; AGROVOC has the least number of terms among thesauri used. However, in case of using NAL2008 the precision is the highest while the number of terms in NAL2008 is not the largest. In contrast, when using NAL2012, which has most number of terms, the precision was not the highest. Recall and F measure behave as the same as precision, which mean that the highest values of recall and F measure was recorded with NAL2008 and lowest values were recorded with AGROVOC thesaurus.

4.1.2. Results discussion of common matches

This sub-section discusses the results of common matches between different thesauri, as follows:

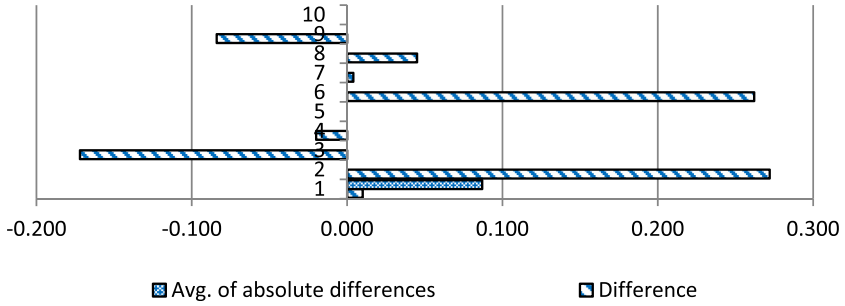

Fig. 12. Differences and average of absolute differences for common matches between NAL2008 and NAL2012.

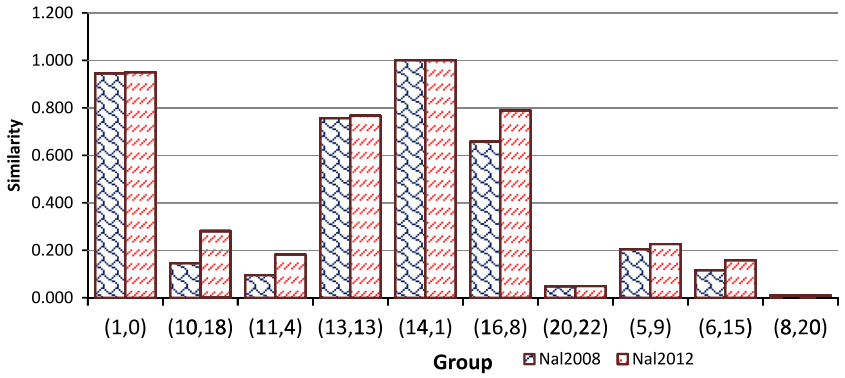

Fig. 13. Average approach versus maximum approach values of common matches between NAL2008 and NAL2012.

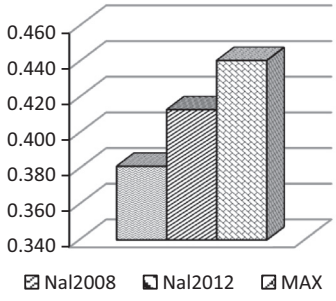

Fig. 14. Enhancement of max approach over average approach of common matches between NAL2008 and NAL2012.

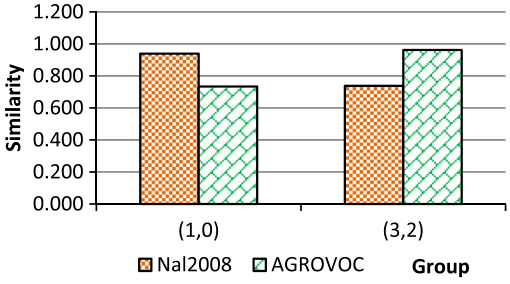

Fig. 15. Similarity values of common matches between NAL2008 and AGROVOC.

4.1.2.1. Results discussion of common matches between NAL2008 and NAL2012 thesauri. Table 5 shows the common matches between results of using NAL2008 Thesaurus and NAL2012 Thesaurus:

From Table 5, it is seen that the Similarity of matches when using NAL2012 Thesaurus was increased or stay constant in 70% of common matches. Common matches between NAL2008 and NAL2012 are more than 40% relative to the number of elements in S_x. Fig. 11 shows the results of using NAL2008 and NAL2012, while Fig. 12 shows the average of absolute differences between similarity values.

It can be seen from Fig. 11 that the similarity when using NAL2012 was equal to or more than the similarity when using NAL2008 in 70% of common matches.

As seen from Fig. 11, Similarity is not increased for all common matches when using the thesaurus with more terms, lead-in terms, and cross relations. As mentioned in Section 2.6, two approaches are used to determine the value of overall similarity for each common group; these approaches are the Average similarity and the Maximum similarity value. It can be seen from Table 5 that the

Table 6
Common matches from results of using NAL2008 thesaurus and AGROVOC thesaurus.

Group no	Pair	Similarity		Absolute value of difference	Similarity avg.	Maximum similarity
		NAL2008 thesaurus	AGROVOC thesaurus			
1	(1,0)	0.939	0.734	0.205	0.837	0.939
2	(3,2)	0.738	0.962	0.224	0.850	0.962
Average		0.839	0.848	0.215	0.843	0.951
Enhancement in similarity of MAX approach relative to avg.		0.112	0.103		0.107	

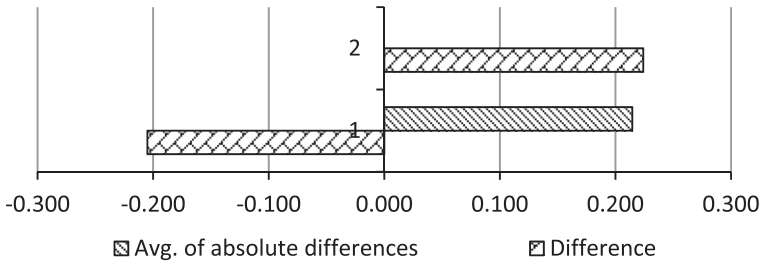

Fig. 16. Differences and average of absolute differences for common matches between NAL2008 and AGROVOC.

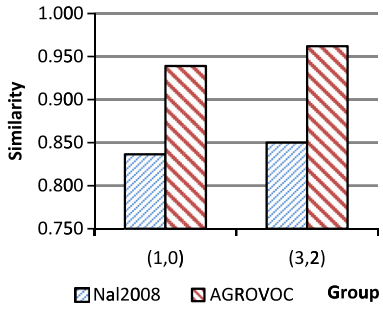

Fig. 17. Average approach versus max approach values of common matches between NAL2008 and AGROVOC.

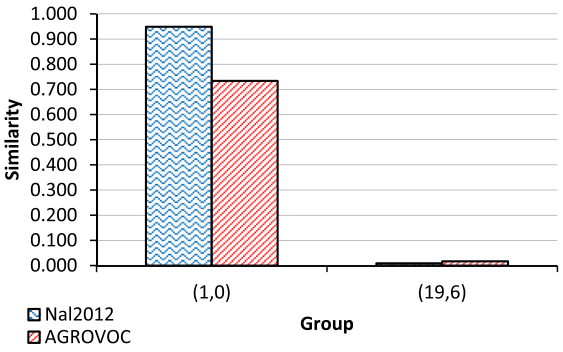

Fig. 19. Similarity values of common matches between NAL2012 and AGROVOC.

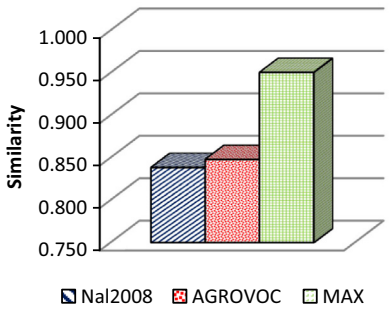

Fig. 18. Enhancement of max approach over average approach of common matches between NAL2008 and AGROVOC.

Maximum approach leads to an enhancement in the average of the similarity by **0.059** and **0.028** for experiment using NAL2008 and NAL2012 consecutively. Fig. 13 shows Average approach versus Maximum approach values, while Fig. 14 shows enhancement of Maximum approach over the Average approach.

4.1.2.2. Results discussion of common matches between NAL2008 and AGROVOC thesauri. Table 6 shows the common matches between results of using NAL2008 thesaurus and AGROVOC thesaurus.

From Table 6 it can be seen that the similarity of matches when using AGROVOC thesaurus, which is the least in terms, lead-in terms, and cross-relations was increased or stay constant in 50% of common matches. Shared matches are about 1% relative to the number of elements in Set 1. Fig. 15 shows the results of using NAL2008 and AGROVOC, while Fig. 16 shows the average of absolute differences between similarity values.

Table 6 shows that the similarity is not decreased for all common matches when using the thesaurus with fewer terms, lead-in terms, and cross relations. Using Max approach enhances the average of the similarity by **0.112** and **0.103** for experiment using NAL2008 and AGROVOC consecutively as shown in Table 6. Fig. 17 shows Average approach versus Max approach values, while Fig. 18 shows enhancement of Max approach over Average approach.

4.1.2.3. Results discussion of common matches between NAL2012 and AGROVOC thesauri. Table 7 shows the common matches between results of using NAL2012 thesaurus and AGROVOC thesaurus.

Table 7 shows that the similarity of matches when using NAL2012 Thesaurus which has more terms, lead-in terms, and cross-relations than AGROVOC, was increased or stay constant in 50% of common matches, common matches are about 1% relative to the number of elements in Set 1. Fig. 19 shows the results of using NAL2012 and AGROVOC, while Fig. 20 shows the average of absolute differences between similarity values:

Table 7
Common matches from results of using NAL2012 thesaurus and AGROVOC thesaurus.

Group no	Pair	Similarity		Absolute value of difference	Similarity avg.	Maximum similarity
		NAL2012 thesaurus	AGROVOC thesaurus			
1	(1,0)	0.949	0.734	0.215	0.842	0.949
2	(19,6)	0.009	0.017	0.008	0.013	0.017
Average		0.479	0.376	0.112	0.427	0.483
Enhancement in Similarity of MAX approach relative to Avg.		0.004	0.108		0.056	

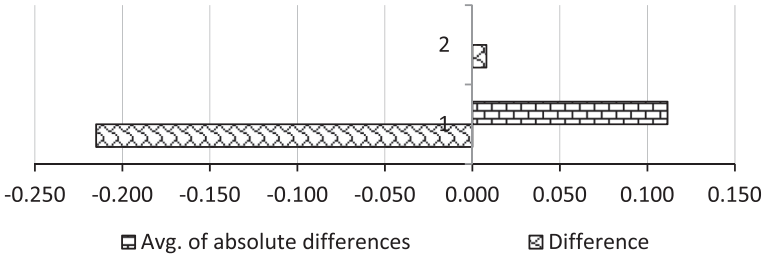

Fig. 20. Differences and average of absolute differences for common matches between NAL2012 and AGROVOC.

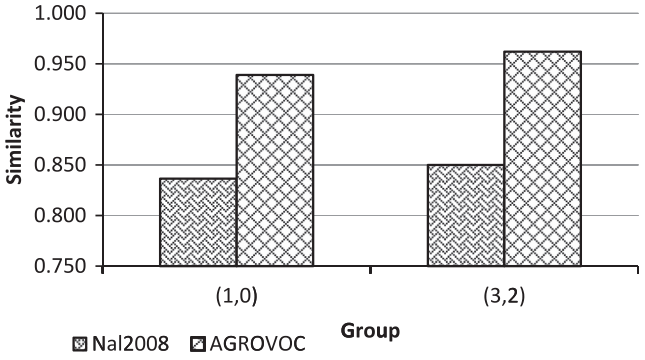

Fig. 21. Average approach versus max approach values of common matches between NAL2012 and AGROVOC.

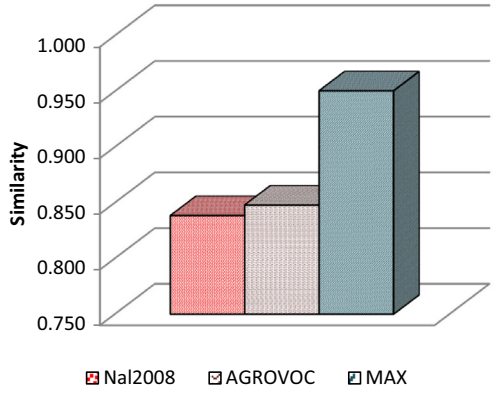

Fig. 22. Enhancement of max approach over average approach of common matches between NAL2012 and AGROVOC.

Table 8
Pair-wise two sided *T*-Test results using common matches.

Thesauri set	*N*	Std.*	Df*	*t*	*p*-Value*
NAL2008–NAL2012	10	0.138	9	−.726	.487
NAL2008–AGROVOC	2	0.303	–	–	–
NAL2012–AGROVOC	2	0.158	–	–	–

Std*: Standard deviation, *p*-values significant at alpha = 0.05, df*: degree of freedom.

Table 9
Similarity means of common matches between AGROVOC and other thesauri.

Thesauri set	*N*	Standard deviation	Similarity avg.
NAL2008–AGROVOC	2	0.303	0.842
NAL2012–AGROVOC	2	0.158	0.247

It seen from Table 7 that the similarity is not decreased for all common matches when using thesaurus with less terms, lead-in terms, and cross relations (AGROVOC). Using Max approach enhances the average of similarity by **0.004** and **0.108** for experiment using NAL2012 and AGROVOC consecutively as shown in Table 7. Fig. 21 shows Average approach versus Max approach values, while Fig. 22 shows enhancement of Max approach over Average approach.

4.1.3. Discussion of significance tests

To evaluate the hypothesis that there is a significant difference between similarities of common matches when using different thesauri, the pair-wise two-sided *T*-Test using common matches among the experiments was performed. Table 8 shows the results of *T*-Test.

It can be seen from the results of *T*-Test that the difference in the similarity of common matches is statistically insignificant for each combination of used thesauri. These insignificant results are due to the small sample size, the limitation of sample size comes from the domain of the experiment. For the pair-wise combinations (NAL2008-AGROVOC and NAL2012-AGROVOC) the statistical *T*-Test is non-applicable because of the too small sample size (2 samples), however it can be seen form Tables 6 and 7 that the similarity average of the common matches between NAL2008 and AGROVOC is too much higher than those between NAL2012 and AGROVOC, as summarized in Table 9.

4.2. Comparison of similarity method calculation

This section presents the comparison between the similarity calculated based on the proposed similarity calculation method (i.e. Eq. (1) which was explained in Section 2.3) and the common cosine similarity measurement. Hence the differences in similarities calculated by every method direct to different final mapping results, because the application of the maximum and second maximum value approach [27]. In the following sub-sections, the similarity of common matches and the overall similarity average are compared and discussed.

4.2.1. Similarity comparison of common matches

To compare the similarity calculated using the proposed method and the cosine similarity, the common matches for each thesaurus were extracted. Fig. 23 shows the comparison.

From Fig. 23, it is seen that the cosine similarity value was higher for all common matches for all thesauri. The reason of this is that the cosine similarity consider the number occurrences of a word (term) in the vector, while the proposed method based on the union operation which eliminate the effect of the repeated words (terms) in the vector and consider each word once. Using cosine similarity in schema matching using thesaurus is leads to higher similarity ratios, however the in automatic schema matching the higher similarity between two elements may cause an incorrect matching since the highly similar elements will be paired

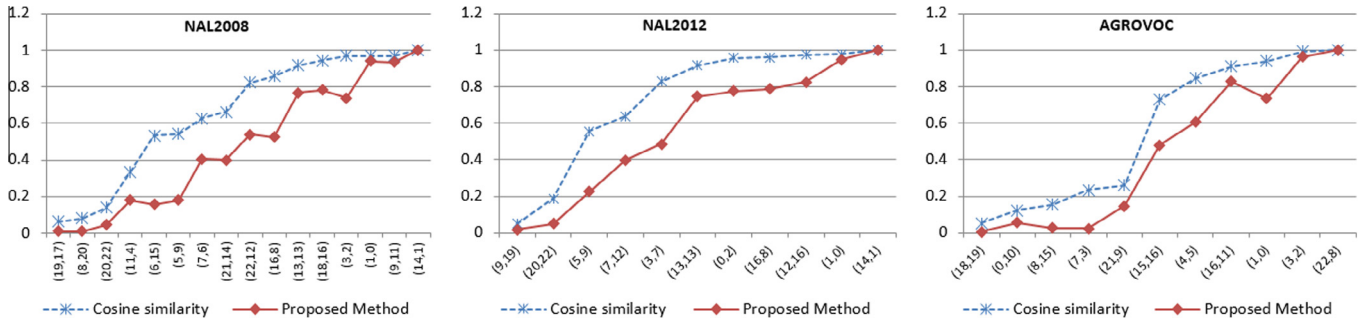

Fig. 23. Similarity comparison of common matches.

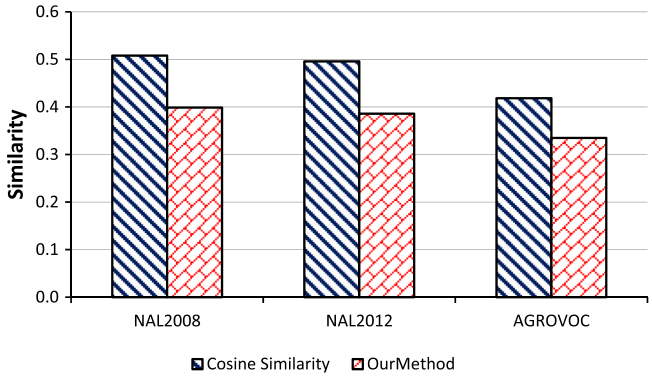

Fig. 24. Overall similarity comparison.

as matching pair, and these elements will not be paired to any other elements. The proposed similarity measurement method as mentioned in Section 2.3 do not consider the occurrences of the term but just the existence.

4.2.2. Overall similarity comparison

The similarity average of final mappings for each thesaurus was compared; Fig. 24 shows that the average of cosine similarities was higher than the average of similarities calculated by the proposed method.

From Fig. 24, it can be seen that the similarity based on cosine method was higher than the similarity based on the equation discussed in Section 2.3. In cosine similarity, the number of occurrences of the term in the vector increases the similarity; however, the proposed method eliminates the effect of multiple occurrences of the term in the vectors, so that the calculated similarity was lower.

5. Conclusion

In this research, thesaurus was utilized to be the core of schema matching process; many experiments were conducted to study the effect of thesaurus size on schema matching quality. Results showed that different mappings were produced because of using different thesauri in the same domain. The common matches between those mappings also have different similarity values. An increment in the average of similarity with distinctive values was recorded. The use of the richest thesaurus (i.e. thesaurus with most number of terms, lead-in terms, and cross relations) does not result the highest precision, recall, and F measure values, whereas the lowest values of precision and recall were recorded when the thesaurus with the least number of terms, lead-in terms, and cross relations was used. The results of schema matching using thesaurus affected with thesaurus size (in aspects of the number of terms and number of cross relations), however the change was statically insignificant. Cosine similarity was also higher than the similarity calculated based on the proposed equation. Predicting the exact value of the change in outcome of schema matching using thesaurus or other thesaurus based applications when using different thesauri to solve the same problem, needs to be deeply studied. However, other factors related to the domain where thesauri are used also affect the results. Currently, we are studying how thesaurus specifications affect the outcome of other IR applications such as document classifiers. The main goal is to generate a mathematical model to predict the quality of the output of IR tools and applications that uses thesaurus as the core of its job, this prediction will depend on thesaurus specifications and domain specifications as parameters.

Acknowledgments

The authors would like to thank our colleagues especially Mr. M. Sirajo, and the people of Software Engineering Research Group (SERG), Universiti Teknologi Malaysia who provided insight and expertise that greatly assisted the research. We also thank Ministry of Science, Technology and Innovation, Malaysia (MOSTI) for the research funding 4S062.

References

[1] K. Golub, Automated subject classification of textual Web pages, based on a controlled vocabulary: challenges and recommendations, New Rev. Hypermedia Multimedia 12 (2006) 11–27.
[2] J.-J. Kuo, H.-C. Wung, C.-J. Lin, H.-H. Chen, Multi-document summarization using informative words and its evaluation with a QA system, in: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, Springer-Verlag, 2002, pp. 391–401.
[3] S. Ralf, H. Johan, S. Stefan, Using thesauri for automatic indexing and for the visualisation of multilingual document collections, in: Ontologies and Lexical Knowledge Bases: Proceedings of the First International OntoLex Workshop, 2000.
[4] R. Steinberger, B. Pouliquen, J. Hagman, Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC, in: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, Springer-Verlag, 2002, pp. 415–424.
[5] T. Sabbah, R. Jayousi, Y. Abuzir, Schema matching using thesaurus, in: Proceeding of 3rd International Conference on Software, Knowledge, Information Management and Applications, 2009, pp. 197–203.
[6] S. Sorrentino, S. Bergamaschi, M. Gawinecki, L. Po, Schema label normalization for improving schema matching, Data Knowl. Eng. 69 (2010) 1254–1273.
[7] L. Po, S. Sorrentino, Automatic generation of probabilistic relationships for improving schema matching, Inform. Syst. 36 (2011) 192–208.
[8] F. Boudin, J.-Y. Nie, M. Dawes, Using a medical thesaurus to predict query difficulty, in: R. Baeza-Yates, A. Vries, H. Zaragoza, B.B. Cambazoglu, V. Murdock, R. Lempel, F. Silvestri (Eds.), Advances in Information Retrieval, Springer, Berlin, Heidelberg, 2012, pp. 480–484.
[9] O. Suominen, C. Mader, Assessing and improving the quality of SKOS vocabularies, J. Data Semantics 3 (2014) 47–73.
[10] C. Batini, C. Cappiello, C. Francalanci, A. Maurino, Methodologies for data quality assessment and improvement, ACM Comput. Surv. 41 (2009) 1–52.
[11] M.A. Hossain, M.A. Ali, M.G. Kibria, M.N. Bhuiyan, A survey of E-commerce of Bangladesh, Int. J. 2 (2) (2013). http://www.ijsr.net/v2i2_02.php.

[12] C. Dong, J. Bailey, A framework for integrating XML transformations, in: D. Embley, A. Olivé, S. Ram (Eds.), Conceptual Modeling – ER 2006, Springer, Berlin, Heidelberg, 2006, pp. 182–195.
[13] A. Gal, Uncertain schema matching: the power of not knowing, in: C. Macdonald, I. Ounis, I. Ruthven (Eds.), CIKM, ACM, 2011, pp. 2615–2616.
[14] J. Gong, R. Cheng, D. Cheung, Efficient management of uncertainty in XML schema matching, VLDB J. 21 (2012) 385–409.
[15] C.C. Aggarwal, in: C.C. Aggarwal (Ed.), Uncertainty in Data Integration Managing and Mining Uncertain Data, Springer, US, 2009, pp. 1–36.
[16] J. Madhavan, P.A. Bernstein, E. Rahm, Generic schema matching with cupid, in: Proceedings of the 27th International Conference on Very Large Data Bases, Morgan Kaufman Publishers Inc., 2001, pp. 49–58.
[17] A. Doan, P. Domingos, A. Halevy, Learning to match the schemas of data sources: a multistrategy approach, Mach. Learn. 50 (2003) 279–301.
[18] J. Madhavan, P.A. Bernstein, A. Doan, A. Halevy, Corpus-based schema matching, in: Proceedings of the 21st International Conference on Data Engineering, IEEE Computer Society, 2005, pp. 57–68.
[19] E. Rahm, P.A. Bernstein, A survey of approaches to automatic schema matching, VLDB J. 10 (2001) 334–350.
[20] P. Shvaiko, J. Euzenat, A survey of schema-based matching approaches, J. Data Semantics IV (2005) 146–171.
[21] L. Zamboulis, XML Schema Matching & XML Data Migration & Integration: A Step Towards the Semantic Web Vision, 2003.
[22] S. Melnik, H. Garcia-Molina, E. Rahm, Similarity flooding: a versatile graph matching algorithm and its application to schema matching, in: Proceedings of the 18th International Conference on Data Engineering, 2002, pp. 117–128.
[23] H.Q. Thang, V.S. Nam, XML schema automatic matching solution, Int. J. Electr., Comput., Syst. Eng. 4 (2010) 68–74.
[24] Princeton University, "About WordNet", WordNet.
[25] L. Xu, Source Discovery and Schema Mapping for Data Integration, Brigham Young University, 2003, p. 137.
[26] A. Huang, Similarity measures for text document clustering, in: Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand, 2008, pp. 49–56.
[27] B. Mirza, C. Laurent, S. Joel, MAXSM: A Multi-Heuristic Approach to XML Schema Matching, 2006.
[28] A. Levitin, Introduction to the Design & Analysis of Algorithms, Addison-Wesley, Reading, MA, 2003.
[29] Y. Abuzir, T. Sabbah, First token algorithm for searching compound terms using thesaurus database, J. Comput. Sci. 8 (2012) 61–67.
[30] H.H. Do, S. Melnik, E. Rahm, Comparison of schema matching evaluations, in: Revised Papers from the NODe 2002 Web and Database-Related Workshops on Web, Web-Services, and Database Systems, Springer-Verlag, 2003, pp. 221–237.
[31] T. Sabbah, Using Thesaurus as a Schema Matching Approach at the Element Level, Unpublished, MSc Thesis, Al Quds University, 2009.
[32] R. Cheng, J. Gong, D.W. Cheung, Managing uncertainty of XML schema matching, in: IEEE 26th International Conference on Data Engineering (ICDE), IEEE, 2010, pp. 297–308.
[33] H.-H. Do, E. Rahm, COMA: a system for flexible combination of schema matching approaches, in: Proceedings of the 28th International Conference on Very Large Data Bases, VLDB Endowment, Hong Kong, China, 2002, pp. 610–621.
[34] A. Miles, A Thesaurus Data Model for British Standard 8723, 2006. <http://alimanfoo.wordpress.com/2006/11/>.