Jurnal Teknologi

Full paper

Schema Matching Quality: Thesaurus as the Matcher

Thabit Sabbah*, Ali Selamat

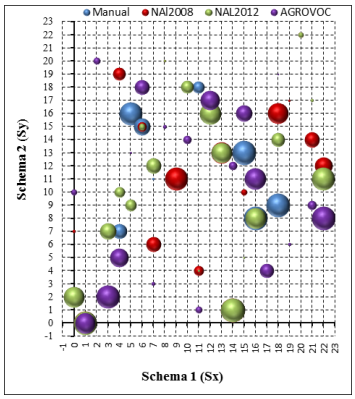
Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

*Corresponding author: sosthabit2@live.utm.my

Article history

Received :1 January 2014
Received in revised form :
1 June 2014
Accepted :10 September 2014

Graphical abstract



Abstract

Thesaurus is used in many Information Retrieval (IR) applications such as data integration, data warehousing, semantic query processing and classifiers. It was also utilized to solve the problem of schema matching. Considering the fact of existence of many thesauri for a certain area of knowledge, the quality of schema matching results when using different thesauri in the same field is not predictable. In this paper, we propose a methodology to study the performance of the thesaurus in solving schema matching. The paper also presents results of experiments using different thesauri. Precision, recall, F-measure, and similarity average were calculated to show that the quality of matching changed according to the used thesaurus.

Keywords: Schema matching; thesaurus; information retrieval; performance

Abstrak

Thesaurus digunakan dalam banyak aplikasi capaian semula maklumat seperti integrasi data, gudang data, pemprosesan soalan semantik dan pengkelasan. Ia juga diguna pakai untuk menyelesaikan masalah padanan skema. Memandangkan banyak thesauri dihasilkan bagi sesetengah bidang, kualiti bagi hasil padanan skema tidak dapat dijangka terutamanya apabila menggunakan thesauri berbeza bagi bidang yang sama. Untuk kertas ini, satu kaedah dicadangkan bagi mengkaji prestasi thesaurus semasa menyelesaikan padanan skema. Kertas ini juga membentangkan hasil bagi eksperimen yang menggunakan thesauri berbeza. Ketepatan, panggilan semula, pengukuran-F, dan purata persamaan dikira untuk menunjukkan bahawa kualiti padanan berubah mengikut thesaurus yang digunakan.

Kata kunci: Skema matching; thesaurus; semula maklumat; pencapaian

© 2014 Penerbit UTM Press. All rights reserved.

1.0 INTRODUCTION

Thesaurus is "a controlled vocabulary arranged in a known order and structured so that the various relationships among terms are displayed clearly and identified by standardized relationship indicators. Relationship indicators should be employed reciprocally" [1]. Information retrieval is one of the notable applications of thesaurus since the first machine-readable thesaurus was published 1950 based on Peter Roget thesaurus [2]. By passing of time, purposes of the thesaurus expanded to include aiding in the general understanding of the subject area, providing 'semantic map' of the subject by showing the inter-relations between concepts, and helping to provide definitions of terms [3].

For more than two decades, thesaurus was put into service in many IR applications. It was used in web document classification [4], summarization [5], indexing [6], and in calculating the semantic similarity of documents written in the same or in different languages [7]. Recently, the advantage of the thesaurus was taken to predict query difficulty in the medical domain [8]. 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]. The use and impact of the thesaurus was not studied widely, because it was assumed that there are no general thesauri with sufficient coverage are available [8]. However, in some particular domains, we can find a high-quality thesaurus; moreover, we can find many thesauri with different coverage abilities in the same domain. Thesaurus was also proposed to solve the problem of schema matching [9].

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. This is because schema matching is one of the basic phases [10] in many applications such as data integration, data warehousing, and semantic query processing. Many approaches and tools were proposed to find the equivalent elements between two schemas such as Cupid [11], and LSD [12], Corpus [13]. In addition, many surveys classifications [14],[15] were published. The earlier use of the thesaurus in schema matching field was restricted to applying thesaurus on the names of schemas' elements [11]. Recently, thesaurus was used to analyze the textual description of schemas' elements as the main step in finding the mapping between schemas [9].

In this paper, we propose a methodology to inspect the impact of the thesaurus on the results of schema matching. In the

rest of this paper, Section 2.0 summarizes most common schema matching approaches. Section 3.0 explains our methodology, and then Section 4.0 shows out experiments. Section 5.0 shows the initial results, as well as a discussion on these results. Finally, we conclude this work in Section 6.0.

2.0 RELATED WORK

During the past few decades, many approaches were proposed to carry out the process of schema matching automatically. Few features of matching process were not in the focus of these approaches. Aspects such as structural, element, linguistics, and data model were discussed widely. In this section, we present a summarization of techniques used in schema matching approaches. Many techniques were employed to carry out matching process; Machine-learning techniques were used in [12]; learner-based approaches contain learner modules and specific module to direct learners. These approaches use neural networks' advantages to find out the similarity between data sources. In [16] the characteristics of object-oriented were exploited to determine the mapping between data sources' attributes. The proposed solution in [16] did not solve the problem as well many other works that use metadata; however, the problem is shifted into another problem that is the ontology mapping problem. Most of the current schema matching tools use rules to perform the matching. Moreover, information such as elements names' and descriptions', data types, hierarchy structure, and constraints are employed in determining the similarity at either element level or schema level [11], [16], [17].

Most effective rule-based schema matching methods usually consist of three phases; linguistic, constraint-based, and structural matching [18]. In the 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 [19] to consider the semantic relationships between elements' names. However, it is common that algorithms in this category use combined methods to get high computed similarity, for example, Cupid [11] matcher exploits linguistic matching in a comprehensively and efficiently manner to produce high similarity [11]. The Incorrect results obtained from linguistic matching phase are usually adjusted in the constraint-based matching phase. Data type constraint, data types' compatibility measurement methods are usually used as the initial solution of incorrect or ambiguous results of linguistic matching phase [20], [11]. The structural matching phase is used to solve the problems of context similarity. These problems generally appear in XML schema matching where the structure document, and the constraints on nodes and edges differs from rational schemas, [18] describes such problems in details.

3.0 METHODOLOGY

This paper proposes a methodology based on thesaurus to carry out the process of schema matching.

Figure 1 shows our framework.

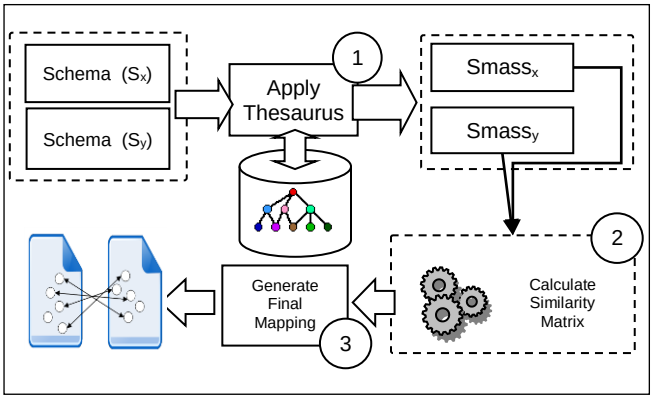


Figure 1 Methodology framework

As shown in the framework, thesaurus is used in solving the problem of schema matching at the element level based on textual analysis of schemas elements' descriptions. Main steps in this framework are as explained below.

First, thesaurus was applied on fields' descriptions, one by one for both schemas *Sx* and *Sy*. Applying thesaurus requires searching for every word from the text into the thesaurus database to get the related terms, and to build up mass of terms related to the term under processing. This process is performed for every word in every element's description in both schemas. Different masses are then collected on the element level into one mass that represents the Result of Applying Thesaurus on element of schema (RAT*ei*S*x*).

Figure 2 shows the algorithm used in applying thesaurus on elements' textual descriptions.

Algorithm 1: Applying Thesaurus on Element's Description

Input:

$S_x = \{(e,desc)_{x1},\ldots,(e,desc)_{xn}\}$
$S_y = \{(e,desc)_{y1},\ldots,(e,desc)_{ym}\}$

for ($S_k \in \{ S_x , S_y \}$) **loop**
  $term\_mass_{jk} \leftarrow \{\}$
  for $((e,desc)_{jk} \in S_k)$ **loop**
    $term\_mass_{jk} \leftarrow \cup$ ***get_term_mass***$(desc_{jk})$

Output:

$Smass_x = \{(e,term\_mass)_{x1},\ldots,(e,term\_mass)_{xn}\}$
$Smass_y = \{(e,term\_mass)_{y1},\ldots,(e,term\_mass)_{ym}\}$

Figure 2 Applying thesaurus on element description algorithm

Then, we calculate the similarity between RAT of each element from *Sx* with all elements from *Sy*, to construct the similarity matrix. Algorithm used in calculating similarity matrix is shown in Figure 3.

Algorithm 2: Calculating Similarity Matrix

Input:
Smass$_x$ ={(e,term_mass)$_{x1}$,…,(e,term_mass)$_{xn}$}
Smass$_y$ ={(e,term_mass)$_{y1}$,…,(e,term_mass)$_{ym}$}
Begin
// create and initialize similarity matrix
SimMatrix←Matrix[n][m]
Initialize SimMatrix; // set all cells to 0
for (e_i ∈ Smass$_x$)
for (e_j ∈ Smass$_y$)
SimMatrix$_{ij}$ ← ***Similarity***(e_i,e_j)
Loop
Loop
Output:
SimilarityMatrix[n][m]

Figure 3 Calculating similarity matrix

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

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

where RAT(e_iS_x) is Result of Applying Thesaurus on the i^{th} element of schema S_x. 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.

Finally, we generate the final mapping set by applying the "Generate Final Mapping" process that implements the maximum and second maximum value algorithm [21]. In this algorithm, a matching (mapping) between the element in the row header and the column header element is considered if the value in their cross cell is the maximum value in the similarity matrix. Then all remaining values in the row and the column are set to zero. This process is repeated until all values in the similarity matrix become zeros or less than the threshold value. The problem of this criterion will arise 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. Figure 4 shows the algorithm used in generating the final mapping from the similarity matrix.

To evaluate the performance of thesaurus in our initial experiments, we use the common measures used in schema matching approaches; precision, recall, F measure, and overall. In the context of information retrieval, precision is used to measure the "exactness" of the solution, while the "completeness" is measured with recall; finally, the F-measure is the weighted harmonic mean that combines the precision and recall into a single unit of measurement.

Algorithm 3: Calculating Similarity Matrix

Input:
S = SimilarityMatrix[n][m]
Variables:
cellIndex=[row,col]
finalMaping={}// set of cell Indexes
Begin
While S contains value > 0
//get the maximum value in the matrix
max ← getMaxValue(S)
// get the x,y index of max value in the matrix
cellIndex(row,col) ← getRowCol(max)
// check for uniqueness
If (max is unique)
// append cell index to the final Mapping list
FinalMapping ← U cellIndex(row,col)
// set similarity value to zeros in row and column
S[row] ← 0
S[col] ← 0
// set max value to negative in the similarity matrix
S[row,col] ← -1 * max
Else
// set all cells equals to max to zero
∀ (S[row,col] = max) : S[row,max] ← 0
End If
Loop
Output:
finalMaping

Figure 4 Generating final mapping algorithm

To calculate precision, recall, and F-measure; we considered the manual matches generated by the domain expert as in [22], and then for each experiment we determined the set of true positives (TP), false positives (FP), and false negatives (FN). TP, FP, and FN are defined as follows:

- TP: Manual matches correctly identified as matches
- FP: Manual matches, detected by the automatic matcher as non-matches.
- FN: Matches detected by the automatic matcher while it is not included in the manual matching.

Based on these sets the quality measures were calculated as follows:

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

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

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

4.0 EXPERIMENTS

To carry out our experiments, we installed Oracle database with Java application developed especially for that purpose. Two sets of courses offered by two different universities in the domain of agriculture were tested to find the equivalent courses between them. For experimental uses, sets were named as follows:

Set One: S_x = (0,1,2,..,22), and
Set Two: S_y= (0,1,2,..,22),

Using different thesauri in the same domain (Agriculture Domain), courses' descriptions in both sets were processed and analyzed. First, we used the National Agricultural Library[1] Thesaurus 2008 Edition (NAL2008), then we used the National Agricultural Library Thesaurus 2012 Edition (NAL2012), and finally, we used the thesaurus presented by Food and Agriculture Organization of the United Nations (AGROVOC Thesaurus[2]). All thesauri were downloaded from the Internet, and processed by special tools to meet our environment; Table 1 shows main specifications of used thesauri.

Table 1 Thesauri specifications

	NAL 2012 Thesaurus	NAL 2008 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

From Table 1 we can see that NAL2012 is the thesaurus that contains the most number of terms, lead-in terms, and cross-relations, while AGROVOC has the least number of the all specifications.

5.0 RESULTS AND DISCUSSION

Two sets of courses used in our experiments were manually matched by an expert [22], results of manual and automatic matching from the experiments based on different thesauri are shown in Figure 5.

In Figure 5, the numbers on x-axis and y-axis represents the elements in schemas, while the bubbles represent the matches between elements. For example we have a matching between element 5 from schema 1 (x-axis) and element 16 from schema 2 (y-axis) in manual matching, this match case is referred as pair (5,16), the size of bubble represents the value of the similarity between each pair of 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.

5.1 Precision, Recall and F-Measure Results

The precision, recall and F measure for each experiment was calculated relative to manual matches. Table 2 shows the values of precision, recall, and F measure.

Table 2 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

From Table 2, it can be seen that the use of rich thesaurus (NAL2012) does not lead to higher precision and recall results. However, the use of AGROVOC thesaurus that has fewer terms, leading terms, and cross-relations cause a low precision and recall values.

Figure **6**Figure 6 shows the relation between precision, recall, and F measure related to the number of terms in each thesaurus.

1 http://agclass.nal.usda.gov/
2 http://aims.fao.org/standards/agrovoc/about

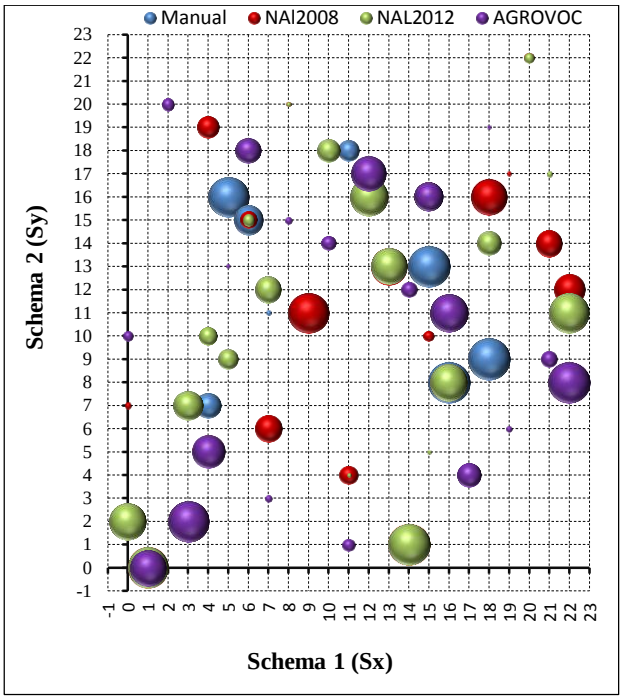


Figure 5 Results of manual and automatic matching

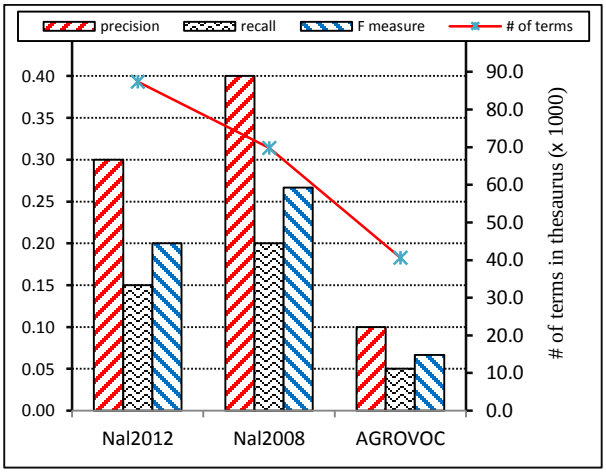


Figure 6 Precision, recall, and f-measure for different thesauri

As seen from Figure 6, the precision, recall and f-measure 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, recall, and f-measure was the greater while the number of terms in NAL2008 is not the highest among all. In contrast, when using NAL2012 which has most number of terms, the precision, recall, and f-measure was not the higher.

5.2 Common Matches Results

In addition, we look to the results from the common matches' view. Common matched are the pairs of nodes that defined as matches by the automatic matcher based on different thesauri. For example, when using NAL2008 the pair (10,18) is defined as a match with similarity value equals to 0.009, also it was defined as a match when using NAL2012 with similarity value equals 0.28.

Table 3 shows the common matches between results of using NAL 2008 Thesaurus and NAL 2012 Thesaurus.

Table 3 Common Matches from Results of Using NAL2008 Thesaurus and NAL2012 Thesaurus

Pair	Similarity (Normalized)	
	NAL 2008 Thesaurus	NAL 2012 Thesaurus
(1,0)	0.939	0.949
(10,18)	0.009	0.281
(11,4)	0.181	0.009
(13,13)	0.766	0.746
(14,1)	1.000	1.000
(16,8)	0.526	0.788
(20,22)	0.045	0.049
(5,9)	0.181	0.226
(6,15)	0.157	0.073
(8,20)	0.009	0.009
Average	**0.381**	**0.413**

As shown in Table 3, the similarity of matches when using NAL 2012 Thesaurus, which has more terms, lead-in terms, and cross-relations than NAL2008, 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. Moreover, the similarity based on the using NAL2012 was equal to or more than the similarity of NAL2008 in 70% of common matches.

In addition, it can be seen that the similarity is not increased for all common matches when we use a thesaurus with more terms, lead-in terms, and cross relations. However, the average similarity is increased from **0.381** to **0.413** for experiment using NAL2008 and NAL2012 consecutively. The increment in similarity for individual pairs and the average similarity for all pairs, reflects the enhancement and increment in terms, lead-in terms, and cross relations between terms between the old edition of the used thesaurus (NAL2008) and the new edition (NAL2012). The relation between these two parameters will be in the focus of our future research.

To evaluate the hypothesis that there is a significant difference between similarities of common matches when using different thesauri, we performed the pair-wise two-sided T-Test using common matches among our experiments. Table 4 shows the results of T-Test:

Table 4 Pair-wise two sided t-test results using common matches

	N	Avg. Similarity	Std.*	df*	t	p-value*
Nal2008-Nal2012	10	0.397	0.138	9	-.726	.487
Nal2008-AGROVOC	2	0.843	0.303	1	-.044	.972
Nal2012-AGROVOC	2	0.427	0.158	1	.928	.524

The results of T-Test in Table 4, shows that the differences in similarity of common matches is statistically insignificant for different thesauri combinations. The insignificant results are referred to the small sample size of sample size, which comes from the limitation of the domain.

6.0 CONCLUSION

It clearly noticed that using different thesauri in the same domain produces distinctive mappings accordingly. The use of a thesaurus with more terms, lead-in terms, and cross relations did not lead to the high precision and recall values. However, the lowest values of precision and recall were recorded at using the thesaurus with the least number of terms, leading terms, and cross relations. The results of schema matching using thesaurus were affected directly with the thesaurus size. From the view of common matches between mappings, an increment in the average of similarity with distinctive values when using different thesauri was recorded. However, the increment is not related to the specifications of the used thesaurus.

Our future work will focus on studying the significance of variation of similarity averages within common matches generated by using different thesauri, and the relation between the enhancement of thesaurus editions and the increment in similarity and average similarity. Moreover, the future study will be extended to include the effect of thesaurus size on the outcome of other tools and applications in IR domain such as document classifiers.

Acknowledgement

The Universiti Teknologi Malaysia (UTM) is acknowledged for some of the facilities utilized during the course of this research and Ministry of Higher Education (MOHE) Malaysia, under research grant R.J130000.7828.4F087 and Ministry of Science Technology and Innovation (MOSTI) under research grant R.J130000.7909.4S062 is acknowledged for the research funding.

References

[1] *American National Standards Institute*. 2005. ANSI/NISO Z39.19-2005.
[2] Masterman, M. 1957. The Thesaurus in Syntax and Semantics. *Mechanical Translation*. 4(1–2): 35–43.
[3] Aitchison, J., D. Bawden, and A. Gilchrist. 1997. *Thesaurus Construction and Use: A Practical Manual*. 3rd ed.
[4] Golub, K. 2006. Automated Subject Classification of Textual Web Pages, Based on a Controlled Vocabulary: Challenges and Recommendations. *New Review of Hypermedia and Multimedia*. 12(1): 11–27.
[5] Kuo, J.-J., *et al*. 2002. 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. 391–401.
[6] Ralf, S., H. Johan, and S. Stefan. 2000. 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*.
[7] Steinberger, R., B. Pouliquen, and J. Hagman. 2002. 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. 415–424.
[8] Boudin, F., J.-Y. Nie, and M. Dawes. 2012. *Using a Medical Thesaurus to Predict Query Difficulty*. In *Advances in Information Retrieval*. R. Baeza-Yates, *et al*. Editors. Springer Berlin Heidelberg. 480–484.
[9] Sabbah, T., R. Jayousi, and Y. Abuzir. 2009. Schema Matching Using Thesaurus. In *Proceeding of 3rd International Conference on Software, Knowledge, Information Management and Applications*.
[10] Dong, C. and J. Bailey. 2006. *A Framework for Integrating XML Transformations*. In *Conceptual Modeling-ER 2006*. D. Embley, A. Olivé, and S. Ram, Editors. Springer Berlin Heidelberg. 182–195.
[11] Madhavan, J., P. A. Bernstein, and E. Rahm. 2001. *Generic Schema Matching with Cupid*. In *Proceedings of the 27th International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc. 49–58.
[12] Doan, A., P. Domingos, and A. Halevy. 2003. Learning to Match the Schemas of Data Sources: A Multistrategy Approach. *Mach. Learn.* 50(3): 279–301.
[13] Madhavan, J., *et al*. 2005. Corpus-Based Schema Matching. In *Proceedings of the 21st International Conference on Data Engineering* IEEE Computer Society. 57–68.
[14] Rahm, E. and P. A. Bernstein. 2001. A Survey of Approaches to Automatic Schema Matching. *The VLDB Journal*. 10(4): 334–350.

[15] Shvaiko, P. and J. Euzenat. 2005. A Survey of Schema-based Matching Approaches. *Journal on Data Semantics*. IV: 146–171.
[16] Zamboulis, L. 2003. *XML Schema Matching & XML Data Migration & Integration: A Step Towards The Semantic Web Vision*,
[17] Melnik, S., H. Garcia-Molina, and E. Rahm. 2002. Similarity Flooding: A Versatile Graph Matching Algorithm and Its Application to Schema Matching. In *Data Engineering, 2002. Proceedings. 18th International Conference on*.
[18] Thang, H. Q. and V. S. Nam. 2010. XML Schema Automatic Matching Solution. *International Journal of Electrical, Computer, and Systems Engineering*. 4(1): 68–74.
[19] Princeton University. 2010. *About WordNet. WordNet.* Princeton University.
[20] Xu, L. 2003. *Source Discovery and Schema Mapping for Data Integration*. Brigham Young University. 137.
[21] Mirza, B., C. Laurent, and S. Joel. 2006. *MAXSM: A Multi-Heuristic Approach to XML Schema Matching*.
[22] Sabbah, T. 2009. *Using Thesaurus as a Schema Matching Approach at the Element Level.* Unpublised MSc. Thesis. Al Quds University.