UNIVERSITI PUTRA MALAYSIA

ENHANCED NORMALIZATION APPROACH TO ADDRESS STOP-WORD COMPLEXITY IN COMPOUND-WORD SCHEMA LABELS

JAFREEN HOSSAIN

FSKTM 2014 26

ENHANCED NORMALIZATION APPROACH TO ADDRESS STOP-WORD COMPLEXITY IN COMPOUND-WORD SCHEMA LABELS

By

JAFREEN HOSSAIN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfillment of the Requirements for the Degree of Master of Science

June, 2014

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirement for the degree of Master of Science

ENHANCED NORMALIZATION APPROACH TO ADDRESS STOP-WORD COMPLEXITY IN COMPOUND-WORD SCHEMA LABELS

By

JAFREEN HOSSAIN

June 2014

Chairman: Nor Fazlida Mohd Sani, PhD

Faculty: Computer Science and Information Technology

An extensive review of the existing research work in the field of schema matching uncovers the significance of semantics in this subject. It is beyond doubt that both structural and semantics aspect of schema matching have been the topic of research for many years and there are strong references available for both. However, an in-depth analysis of all the available approaches suggests there are further scopes for improvement in the field of semantic schema matching. Normalization and lexical annotation methods using WordNet have been proposed in several studies. However the results show comparatively poor accuracy due to the presence of stop-words in schema labels. Stop-words have previously been ignored in most studies resulting in false negative conclusions. This research work proposes, NORMSTOP (NORMalizer of schemata having STOP-words), an improved schema normalization approach, addressing the complexity of stop-words (e.g. 'by', 'at', 'and,' or') in Compound Word (CW) schema labels. NORMSTOP isolates these labels during the preprocessing stage and resets the base-form to a relevant WordNet term, or an annotable compound noun; using a combined set of WordNet features like Attributes, Derivationally Related Forms, and LexNames. When tested on the same real dataset used in the earlier approach - (NORMS or NORMalizer of Schemata), NORMSTOP shows up to 13% improvement in annotation recall measurement. This level of improvement takes the overall schema matching process one step closer to perfect accuracy; and the lack of it exposes a gap in expectation, especially in today's databases where stop-words are in abundance.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia
Sebagai memenuhi keperluan untuk ijazah Master Sains

PENDEKATAN NORMALISASI DIPERTINGKATKAN UNTUK MENANGANI KOMPLEKSITI KATA-HENTI DI DALAM LABEL SKEMA KATA MAJMUK

Oleh

JAFREEN HOSSAIN

Jun 2014

Pengerusi: Nor Fazlida Mohd Sani, PhD

Fakulti: Sains Komputer dan Teknokogi Maklumat

Kajian yang mendalam bagi kerja-kerja penyelidikan yang sedia ada dalam bidang padanan skema mendedahkan kepentingan semantik dalam perkara ini. Ia adalah di luar keraguan bahawa kedua-dua aspek iaitu struktur dan semantik daripada padanan skema telah menjadi topik penyelidikan selama bertahun-tahun dan terdapat rujukan yang kukuh disediakan untuk kedua-duanya. Walau bagaimanapun, analisis yang mendalam daripada semua pendekatan ada menunjukkan terdapat skop lagi untuk penambahbaikan dalam bidang padanan skema semantik. Penormalan dan kaedah anotasi leksikal menggunakan WordNet telah dicadangkan dalam beberapa kajian. Walau bagaimanapun keputusan menunjukkan ketepatan yang kurang baik disebabkan oleh kehadiran kata-henti dalam label skema. Kata-henti sebelum ini telah diabaikan dalam kebanyakan kajian menyebabkan kesimpulan negatif palsu. Penyelidikan ini mencadangkan, NORMSTOP, penambahbaikan pendekatan penormalan skema, menangani kerumitan kata-henti (contohnya 'oleh ', ' di ', ' dan, ' atau ') dalam label skema kata majmuk. NORMSTOP mengasingkan label ini semasa peringkat pra-pemprosesan dan mengeset semula bentuk asas untuk pemetaan istilah WordNet, atau anotasi kata nama; menggunakan set gabungan ciri-ciri WordNet seperti Atribut, Bentuk Terbitan Berkaitan, dan LexNames. Apabila diuji pada dataset sebenar yang sama digunakan dalam pendekatan yang lebih awal (NORMS), NORMSTOP menunjukkan peningkatan sehingga 13% dalam anotasi pengukuran ingat. Tahap peningkatan mengambil skema proses pemadanan keseluruhan satu langkah lebih dekat dengan ketepatan yang sempurna dan kekurangan itu mendedahkan jurang yang besar dalam jangkaan, terutamanya di dalam pangkalan data hari ini di mana terdapat kata henti yang banyak.

ACKNOWLEDGEMENTS

At the very beginning I would like to express my sincere gratitude to Allah almighty, for helping me all through, and showing me the right path at the right time for accomplishing this Master Thesis. I also would like to thank my late parents whose good wishes are always with me.

Afterwards, I wish to thank my supervisor Dr. Nor Fazlida Mohd Sani for her ever useful guidance, comments, remarks and engagement through the learning process of this master thesis. I owe my deepest gratitude to Dr. Lilly Suriany Affendey for introducing me to the topic of schema matching as well for all the support on the way. Also I would like to thank Dr. Iskandar Ishak for his valuable inputs during all our meetings.

I am indebted to Dr. Khairul Azhar Kasmiran, this thesis would not have been possible without the help, support and patience of his technical support and valuable time.

I would like to acknowledge the financial, academic and technical support of Universiti Putra Malaysia and particularly the awarding of a Postgraduate Research assistantship that provided the necessary financial support for this research.

Lastly, I would like to thank my best partner and friend, my husband Mirza Asif Haider and my kids for the unconditional support and love throughout this relatively long process. I will be grateful forever for your love.

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of Master of Science. The members of the Supervisory Committee were as follows:

Nor Fazlida Mohd Sani, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Lilly Suriani Affendey, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Iskandar Ishak, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

____

BUJANG BIN KIM HUAT, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

Declaration by graduate student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature: ________________________ Date: ___________________

Name and Matric No.: Jafreen Hossain, GS34731

Declaration by Members of Supervisory Committee

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: ___________________
Name of
Chairman of
Supervisory
Committee: ___________________

Signature:___________________
Name of
Member of
Supervisory
Committee___________________

Signature: ___________________
Name of
Member of
Supervisory
Committee: ___________________

Signature:___________________
Name of
Member of
Supervisory
Committee___________________

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CN	Compound Noun
CW	Compound Word
DRF	Derivationally Related Form
JWI	Java WordNet Interface
NLP	Natural Language Processing
NORMS	NORMalizer of Schemata
NORMSTOP	NORMalizer of schemata having STOP-words
ODBTools	Description Logic Based Tool
OWL	Web Ontology Language
PCT	Probabilistic Common Thesaurus
PWSD	Probabilistic Word Sense Disambiguation
SW	Stop-word
WN	WordNet
WSD	Word Sense Disambiguation

CHAPTER 1

INTRODUCTION

The advancement of information and communication technology has opened doors for many data sources to communicate with each other in a semantic web. At the same time it has created data heterogeneity problems in various application domains. Large amount of data is created every day by different sources in different formats. The value of data increases when it can be linked with other data, thus data integration is a major creator of value. So, data integration and data sharing are getting important for many application domains. But at the same time, the semantic integration is getting crucial and complex due to this large scale data and its heterogeneous nature. This heterogeneity can be in terms of data source format, types, representation, or semantic interpretation.

The schema matching problem is considered by many researchers as one of the bottlenecks for semantic integration. It is not a new research area and has received increasing attention since the 1970s (Islam et al., 2008). Numerous matching approaches, strategies and algorithms have been developed. Schema matching is the task of identifying semantic correspondences between elements of metadata structures such as database schemas, entity relationship diagrams, and ontologies. It is significant for interoperability and data integration in various applications such as data warehousing, integration of web sources, and ontology alignment in the semantic web. In this research study, the main focus is on schema normalization used in semantic schema matching in the context of data integration.

Currently, the schema matching process has improved from fully manual to semi-automatic after years of research by numerous researchers. The process is still not fully automated, has shortcomings in lots of areas, and needs improvements that consider the increasing number of data, schema and data sources. Schemas developed for different application domains can be dissimilar in nature, i.e. although the data is semantically related, the structure and syntax of its representation are different.

Automatic or semi-automatic schema matching has to deal with problems arising from the heterogeneity of data sources which can be distinguished into two main types of heterogeneity: structural and semantic heterogeneity (Sorrentino et al., 2011). Structural heterogeneity means differences among attribute types, formats, or models whereas semantic heterogeneity means differences in the meaning of schema elements. In this research study, we will mainly focus on semantic heterogeneity and its probable solutions.

The main motivation of this research study is the work done by Sorrentino et al. (2011), which focused on schema normalization and lexical annotation methods. It has been proven that schema normalization approaches improve the lexical relationship and matching accuracy among schema labels. Lexical annotation (i.e. annotation with reference to a lexical resource/dictionary, e.g. WordNet) helps to relate a "meaning" to schema labels. However, the accuracy of semi-automatic lexical annotation methods on real life schemas still suffer from the problem of non-dictionary words such as compound words (CWs), abbreviations and acronyms. Schema normalization approaches can help to resolve this problem and increase the number of similar schema labels.

Although different strategies were developed to solve this kind of problem including schema normalization approaches (Sorrentino et al., 2011), there is still room for improvement and future work. Future work might include finding the meaning of different compound words having prepositional-verbs (e.g. "writtenBY"), conjunctions (e.g. "and", "or"), digits or stop-words (e.g. "by", in", "to" etc.) in schema labels. Moreover, future effort might consider the inclusion and integration of other domain-specific resources (such as ontologies, thesauri, glossaries and Wikipedia) to address the problem of the specific domain terms in schema labels (e.g., the biomedical term "aromatase" which is an enzyme involved in the production of estrogen); the use of multi-language lexical resources in order to be able to normalize and annotate schema labels in different languages. Also more work can be done to improve the number of false positive and false negative relationships. Another relevant future research could possibly be the inclusion of instance-based matching techniques to improve the automatic annotation and relationship discovery processes among schema labels.

1.1 Problem Statement

As mentioned by Sorrentino et al. (2011), the weakness of a thesaurus, like WordNet, is that it does not always cover the detail information of a specific domain and domain-dependent terms or words, or non-dictionary words (such as Compound words, abbreviations, acronyms etc). So this kind of non-dictionary words in schema labels strongly affects the automatic lexical annotation technique. To address this problem they presented a method for schema label normalization which expands abbreviations and automatically annotates Compound Nouns (CNs) by enriching WordNet with new meanings.

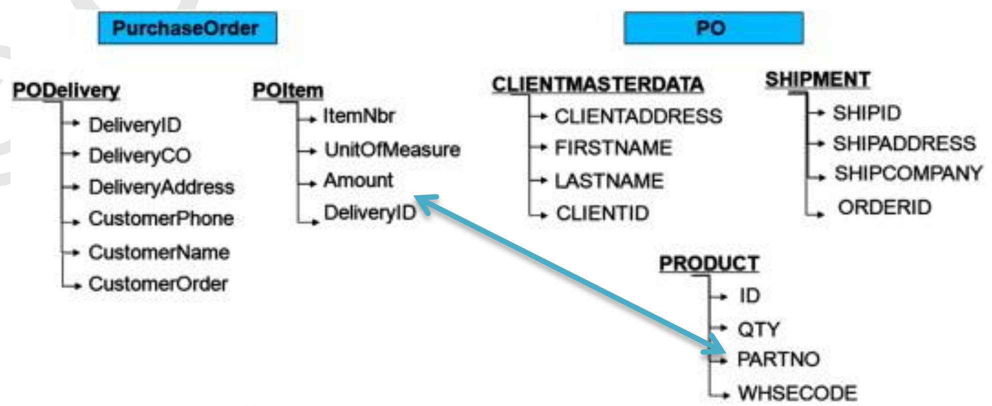

Figure 1: Schema Matching with Labels Having Abbreviation

For example, in Figure 1, “amount” and “QTY” (abbreviation for “quantity”) is the corresponding schema label. So in this kind of case, it is difficult to identify a synonym relationship between the elements “amount” and “QTY” without abbreviation expansion.

With regards to the schema label normalization method, they mentioned some limitation and future improvements in their work which would take into consideration the main problem during the experimental evaluation: The presence of stop-words (e.g. “to”, “at”, “and” etc.) in schema labels; and the problem of false negative (ie. missing right annotation) non-dictionary words during the identification step of schema normalization (Sorrentino et al., 2011).

Po and Sorrentino (2011) also stated the recall rate was affected by the existence of non-endocentric (endocentric CNs is a kind of CNs consisting of a head and modifier) CNs (such as “ManualPublished”, “isMember” or “InProceedings”) in the schemas for all the data sets and that their method could not identify.

So, the limitations summarized from Sorrentino et al. (2011) that were not considered while processing schema label normalization are:

1) Other kinds of multi-word units (e.g. prepositional verbs such as “WrittenBy”)
2) The use of conjunctions (such as “and” or “or”) in schema and ontology labels
3) The presence of stop-words (e.g. “to”, “at”) in schema and ontology labels

Considering the limitations mentioned in Sorrentino et al. (2011), one specific problem has been identified summarizing the three problems mentioned above which needs improvement:

Problem of the presence of stop-words (e.g. “to”, “at”, “and” etc.) in schema labels resulting false negative lexical annotation during schema normalization process (Sorrentino et al., 2011).

1.2 Objective

The objective of the research work is to propose an approach for solving the problem of stop-words in schema labels and improve the lexical annotation of schema label normalization by reducing false negative (ie. missing right annotation) results.

1.3 Delimitation

In our research we will focus only on the Compound Word (CW) annotation which will include Compound Nouns (CNs), or Compound Word formats containing "stop-words" and relevant false negative (missing a right annotation) problem. We will not consider all the "stop-words" used in natural language processing (NLP) since only some common stop-words are used in database designing. The main focus of the research is on stop-words found in the test dataset. Those are "in", "by", "at", "to", "from", "on", "since", "upto", "until", "till", "is", "are", "was", "were", "or" respectively.

1.4 Assumption

In order to fulfill the above mentioned objective, we assume that a fully functional schema normalization tool is implemented and available, in which we can add and run the newly developed algorithm.

1.5 Research Significance

Schema matching is an important and essential process in different domains including e-commerce, data-integration, health-care and many more. By identifying the stop-word in compound word schema labels, the proposed approach would reduce the false negative results in schema normalization and annotation process which is an integral part of schema matching.

1.6 Outline of the Thesis

The rest of the thesis is organized as follows:

Chapter 2 discusses the relevant literature review and some previous approaches on schema and ontology matching. It also details schema normalization approaches and NORMS (NORMalizer of Schemata), an existing tool to perform schema label normalization to enhance the automatic result of schema matching process and some open problems of this area. At the end, it focused on the specific problem of "stop-words" in schema label which is the main focus of this research work. Chapter 3 states the methodology to solve the problem mentioned in chapter 1 and also discusses the new proposed approach "NORMSTOP" and its step by step procedures. Chapter 4 focuses on explaining the implementation of the proposed approach. Chapter 5 details out the evaluation of its results in comparison with previous NORMS approach. Chapter 6 concludes the thesis, mentioning the main contribution and discusses some future opportunities in the same domain.

REFERENCES

Agrawal, D., Bernstein, P. Bertino, E., Davidson, S., Dayal, U., Franklin, M., Gehrke, J., Haas, L., Halevy, A., Han, J., Jagadish, H., V., Labrinidis, A., Madden, S., Papakonstantinou,Y., Patel, J., M., Ramakrishnan, R., Ross, K., Shahabi, C., Suciu, D., Vaithyanathan, S., Widom, J., (2012). *Challenges and opportunities with Big Data*, a community white paper developed by leading researchers across the United States.

Aumueller, D., Do, H. H., Massmann, S., and Rahm, E.(2005). *Schema and ontology matching with COMA++*. In Proc. of Special Interest Group on Managenegment of Data, SIGMOD'05, New York, NY, USA, pages 906–908. ACM.

Banek, M., Vrdoljak, B., Tjoa, A.M. (2007). *Using Ontologies for Measuring Semantic Similarity in Data Warehouse Schema*, In proceeding of: Telecommunications, ConTel 2007. 9th International Conference on, IEEE Xplore.

Banek, M., Vrdoljak, B., Tjoa, A.M., Skocir, Z. (2007). *Automating the Schema Matching Process for Heterogeneous Data Warehouses*. In DaWaK (pp. 45–54).

Bergamaschi, S., Po, L., and Sorrentino, S. (2008). *Automatic annotation for mapping discovery in data integration systems*, in SEBD, S. Gaglio, I. Infantino, and D. Sacca`, Eds., pp. 334–341.

Bergamaschi, S., Castano S, Vincini, M. (1998). MOMIS, An Intelligent System for the Integration of Semi Structured and Structured Data, *INTERDATA*.

Bergamaschi, S., Beneventano, D., Po, L., Sorrentino, S. (2011). Automatic Normalization and Annotation for Discovering Semantic Mappings, Search Computing II, LNCS 6585, pp. 85–100, *Springe.*

Bernstein P., A., Madhavan, J., Rahm, E. (2011). *Generic Schema Matching, Ten Years Later*, Proceedings of the VLDB Endowment, Vol 4, No.11.

Bollegala, D., Honma, T., Matsuo, Y., Ishizuka, M. (2008). *Mining for personal name aliases on the web*. In WWW (pp. 1107–1108).

Budanitsky, A., Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1), 13–47.

Cate, B. T., Dalmau, V., Kolaitis, P. G. (2012). *Learning schema mappings*. Proceedings of the 15th International Conference on Database Theory, , Berlin, Germany.

Chena, N., Heb, J., Yanga, C., Wanga, C. (2012). A node semantic similarity schema-matching method for multi-version Web Coverage Service retrieval, *International Journal of Geographical Information Science.*

Chiticariu, L., Kolaitis, P.G., and Popa, L. (2008). Interactive generation of integrated schemas. In [Wang, 2008], pages 833-846.

Cristian, (2008), Free Stop-Word List in 23 languages, http://www.semantikoz.com/blog/free-stop-word-lists-in-23-languages/

Do, H. H, Rahm E. (2002). COMA - *A system for flexible combination of schema matching approaches*, Proceedings of the 28th VLDB Conference, Hong Kong, China.

Dragut, E., Fang, F., Sistla, P., Yu., C., Meng, W., (2009). Stop Word and Related Problems in Web Interface Integration, *VLDB Endowment*, *ACM.*

Feng, Y., Zhao L., Yang J. (2010). GATuner: Tuning Schema Matching Systems using Genetic Algorithms, *IEEE.*

Finlayson, Alan, M. (2014). *Java Libraries for Accessing the Princeton Wordnet: Comparison and Evaluation.* Proceedings of the 7th Global Wordnet Conference. Tartu, Estonia.

Fox, C., J. (1992). *Lexical Analysis and Stoplists*, Information Retrieval: Data Structures & Algorithms.

Gal, A., Sagi, T., Levy, E., Miklos, Z., (2012). Making Sense of Top-k Matching, *IIWeb*, Scottsdale, AZ, USA.

Gillani, S., Naeem, M., Habibullah, R., Qayyum, A., (2013). Semantic Schema Matching Using Dbpedia, I.J. *Intelligent Systems and Application.*

He, B., Chang, K., C., C. (2004). *A holistic paradigm for large scale schema matching*. SIGMOD Rec., 33(4):20–25.

Hill, E., Fry, Z. P., Boyd, H., Sridhara, G., Novikova, Y., Pollock, L, L., Vijay-Shanker, K. (2008). AMAP: automatically mining abbreviation expansions in programs to enhance software *ACM Trans. Knowl. Discov. Data* maintenance tool.

Hlaing, S. (2009). Ontology based Schema Matching and Mapping Approach for Structured Databases.ICIS, November 24-26, Seoul, Korea.

Islam, A., Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity, *ACM Transactions on Knowledge Discovery from Data*, Vol.2, No.2, Article 10.

Islam, A., Inkpen, D. Z., Kiringa, I. (2008). Applications of corpus-based semantic similarity and word segmentation to database schema matching. *VLDB J.*, 17(5):1293–1320.

Jian, N., Hu, W., Cheng, G., Qu. Y. (2010). *FalconAO: Aligning Ontologies with Falcon*, Department of Computer Science and Engineering Southeast University.

Gong, J., Cheng, R., Cheung, D., W. (2012). Efficient management of uncertainty in XML schema matching, The VLDB Journal — The International Journal on Very Large Data Bases, v.21 n.3, p.385-409.

Karasneh, Y., Ibrahim, H., Othman, M., Yaakob, R. (2010). *Challenges in Matching Heterogeneous Relational Databases Schemas*, IKE'10 - 9th International Conference on Information and Knowledge Engineering – USA.

Kavitha, C., Sadasivam, G. Sudha, S., Shenoy, Sangeetha, N. (2011). Ontology Based Semantic Integration of Heterogeneous Databases, *European Journal of Scientific Research,* ISSN 1450-216X, Vol.64 No.1, pp. 115-122.

Kawahara, M., Kawano, H. (2001). Mining Association Algorithm with Improved Threshold Based on ROC Analysis, IEEE.

Kim, B., Namkoong, H., Lee, D., Hyun, S. J. (2011). *A Clustering Based Schema Matching Scheme for Improving Matching Correctness of Web Service Interfaces*, International Conference on Services Computing, IEEE.

Kopcke, H., Thor, A., Rahm, E. (2010). *Evaluation of entity resolution approaches on real-world match problems*. PVLDB, 3(1), 484–493.

Lee, C. Y., Ibrahim, H., Othman, M., Yaakob, R. (2009). *Reconciling Semantic Conflicts in Electronic Patient Data Exchange*, Proceedings of iiWAS, Kuala Lumpur Malaysia.

Lee, Y., Sayyadian, M., Doan, A., Rosenthal, A., S. (2007). eTuner: tuning schema matching software using synthetic scenarios. *The VLDB Journal*.

Levi, J.N. 1978.*The Syntax and Semantics of Complex Nominals*. New York: Academic Press.

Li, J., Wang, G. A., Chen, H. (2011). Identity matching using personal and social identity features. *IEEE Transactions on Knowledge and Data Engineering identity features*. Information Systems Frontiers, 13(1), 101–113.

Li, J., Tang, J., Li, Y., Luo, Q. (2009). RiMOM: A Dynamic Multistrategy Ontology Alignment Framework. IEEE Trans. Knowl. Data Eng. 21(8), 1218-1232.

Li, Y., Bandar, A., McLean, D. (2003). An approach for Measuring Semantic Similarity between Words Using Multiple Information Sources., 15(4), 871–882.

Liu, X., Zhou, M., Wei, F., Fu, Z., Zhou, X. (2012). *Joint Inference of Named Entity Recognition and Normalization for Tweets*, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 526–535, Jeju, Republic of Korea, Association for Computational Linguistics.

Madhavan, J., Bernstein, P. A., Rahm, E. (2001). Generic Schema Matching with Cupid. In Apers, P. M. G., Atzeni, P., Ceri, S., Paraboschi, S., Ramamohanarao, K., and Snodgrass, R. T., editors, Proc. of the 27th International Conference on Very Large Data Bases (VLDB 2001), September 11-14, 2001, Roma, Italy, pages 49–58. Morgan Kaufmann.

Mandi. (2009). List of English Stop Words. http://norm.al/2009/04/14/list-of-english-stop-words/

Martinez-Gil, J., Aldana-Montes, J. F. (2013). Semantic similarity measurement using historical Google search patterns, *Information Systems Frontiers, Springer Link*.

Miller, R. J., Fisla, D., Huang, M., Kalmuk, D., Ku, F., and Lee, V. (2001). The Amalgam Schema and Data Integration Test Suite.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*, 3:235–244.

Nastase, V., Sokolova, M., Szpakowicz, S. (2006). Learning Noun-Modifier Semantic Relations with Corpus-based and WordNet-based Features, *American Association for Artificial Intelligence.*

Nguyen, T. H., Nguyen, H., Freire, J. (2010). *PruSM: a prudent schema matching approach for web forms,* Proceedings of the 19th ACM international conference on Information and knowledge management, October 26-30, 2010, Toronto, ON, Canada.

Partyka, J., Khan, L., Thuraisingham, B. (2009). *Semantic Schema Matching Without Shared Instances*, International Conference on Semantic Computing, IEEE.

Po, L., Sorrentino, S. (2011). Automatic generation of probabilistic relationships for improving schema matching. *Information Systems Journal*, Special Issue on Semantic Integration of Data, Multimedia, and Services, 36(2):192208.

Ramasubramanian, C., Ramya, R. (2013). Effective Pre-Processing Activities in Text Mining. International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013.

Rahm, E., Bernstein, P. A. (2001). A survey of approaches to automatic schema matching, *The VLDBJournal* 10: 334–350.

Rahm E, Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. University of Leipzig, Germany.

Retzer, S., Yoong, P., Hooper, V. (2012). *Inter-organisational knowledge transfer in social networks: A definition of intermediate ties. Information Systems Frontiers*, 14(2), 343–361. Matching Process. In CONTEL (pp. 227–234).

Sabbah, T., Jayousi, R., Abuzi, Y. (2009) *Schema Matching Using Thesaurus*. in Proceeding of 3rd International Conference on Software, Knowledge, Information Management and Applications.

Salton, G., McGill. M. J., (1983). *Introduction to modern information retrieval*, McGraw-Hill.

Shvaiko, P., Euzenat, J. (2004). *A classification of schema-based matching approaches*. In Proceedings of the Meaning Coordination and Negotiation Workshop at ISWC04.

Shvaiko, P., Euzenat, J. (2005). A survey of schema-based matching approaches, University of Trento, Povo, Trento, Italy, 3730:146–171.

Shvaiko, P., Giunchiglia, F., Yatskevich, M. (2010). Semantic matching with S-Match. *Semantic iWeb Information Management: a Model-Based Perspective*, XX:183–202.

Sorrentino, S., Bergamaschi, S., Gawinecki, M., Po, L. (2010). Schema label normalization for improving schema matching. *DKE Journal*, 69(12):12541273.

Sorrentino, S., Bergamaschi, S., Gawinecki, M. (2011). *NORMS: an automatic tool to perform schema label normalization.* In Press, Accepted Manuscript (Demo Paper), IEEE International Conference on Data Engineering, ICDE 2011, April 11-16, Hannover.

Wang, T. (2006). SeMap: a generic schema matching system. University of British Columbia.

Xue, W., Pung, H., Palmes, P. P, Gu, T. (2008) Schema matching for context-aware computing. In: Proceedings of the 10th international conference on ubiquitous computing, UbiComp ’08. pp 292–301.

Yang, Y., Chen, M., Gao, B. (2008). *An Effective Content-based Schema Matching Algorithm*, International Seminar on Future Information Technology and Management Engineering, IEEE.

Zhao, C., Shen, D., Kou, Y., Nie, T., Yu, G. (2012). *A Multilayer Method of Schema Matching Based on Semantic and Functional Dependencies*, Ninth Web Information Systems and Applications Conference (WISA).