

Discovering Lightweight Ontologies using the Web

Wilson Wong

*School of Computer Science & Software Engineering
The University of Western Australia
35 Stirling Highway, Crawley
WA 6009, Australia
wilson@csse.uwa.edu.au*

Abstract—The need for knowledge structures such as domain thesauri and ontologies by intelligent applications is becoming more and more pressing. Most research on ontology construction conducted in the academia remains not practical to real-world applications. The use of non-dedicated techniques and the dependence on static resources are among the causes to many problems such as non-portability and high maintenance cost faced by existing systems. This paper reports a work in progress which aims to address such problems. Initial experiments using a working prototype of the system revealed great potentials in constructing lightweight domain ontologies using real-world texts.

I. INTRODUCTION

Knowledge structures such as ontologies are valuable assets for a myriad of intelligent applications ranging from information exploration by individuals to knowledge acquisition in organisations. The four commonly used knowledge organisation structures, namely, *controlled vocabulary*, *taxonomy*, *thesaurus* and *ontologies* are often confused and used interchangeably. The key distinction between these schemes lies in the intricacy of the relations and the overall structures. At one end, a controlled vocabulary is the simplest structure consisting of only a flat list of domain terms. At the other end, an ontology supports complex relations and includes axioms for constraint specifications. With the growing popularity of ontology construction in recent years, more and more researchers are looking into this problem area. While many of such research are making progress in the aspect of term recognition and to a lesser extent, relation acquisition, they are far from being able to construct full-fledged domain ontologies. In fact, the majority of the systems are delivering *lightweight ontologies*, which can either be thesauri or taxonomies.

During the course of our research, we have uncovered two general classes of problems faced by existing systems offering lightweight ontologies. The first general problem is concerned with the use of non-dedicated techniques in ontology construction systems. Since ontology construction is a relatively new multi-disciplinary area, many techniques employed in existing systems were directly borrowed from related fields such as Information Retrieval, Text Mining and Information Extraction. As a result, many of the peculiarities involved in transforming real-world natural language texts to machine-readable domain ontologies were not taken into consideration by existing systems. Secondly, many systems rely on static rules and patterns for identifying relations. Coupled with the

use of scarce domain corpora for statistical analysis, and the dependence on rare domain and linguistic knowledge, many existing systems are not extensible across different application domains.

In this paper, we present a general overview of a work in progress aimed at addressing the two problems discussed above. We proposed an approach consisting of dedicated techniques for constructing lightweight ontologies that extracts and recognises domain terms, and discovers semantic relations between the terms using only dynamic resources on the Web such as Google and Wikipedia. Since the construction of lightweight ontologies is a critical step towards achieving full-fledged ontologies, our proposed approach will prove to be indispensable for future work in ontology learning. In Section II, we provide an overview of our system architecture. In Section III, we demonstrate the process of ontology construction using a working prototype of our system tested on real-world domain texts. We conclude this paper in Section IV with our plans for future work.

II. SYSTEM ARCHITECTURE

In this section, we detail out the architecture of our lightweight ontology construction system. The system is comprised of four main phases as shown in Figure 1 (using shaded rectangles). These phases are *text cleaning*, *text processing*, *term recognition* and *relation acquisition*. Text cleaning removes noise such as spelling errors, abbreviations and improper casings from texts. Text processing then extracts coherent three-part structures from the texts using part of speech and dependency information. Term recognition identifies domain-relevant terms from the list of candidates produced during text processing. During the last phase of relation acquisition, the recognised terms are annotated with semantic relations to construct an ontology.

The specific functionalities required in each phase are further decomposed into modules as depicted in Figure 1 using rounded rectangles. Our novel techniques are shown in the same figure using white rounded rectangles. Our ontology construction system requires two types of corpora. A *contrastive corpus* is constructed by performing web crawling to gather web pages from general sources such as Reuters, Discovery and CNet. Readily available general text corpora such as Reuters-21578 [5], GENIA Corpus [4] and British National Corpus [1] can also contribute to the contrastive

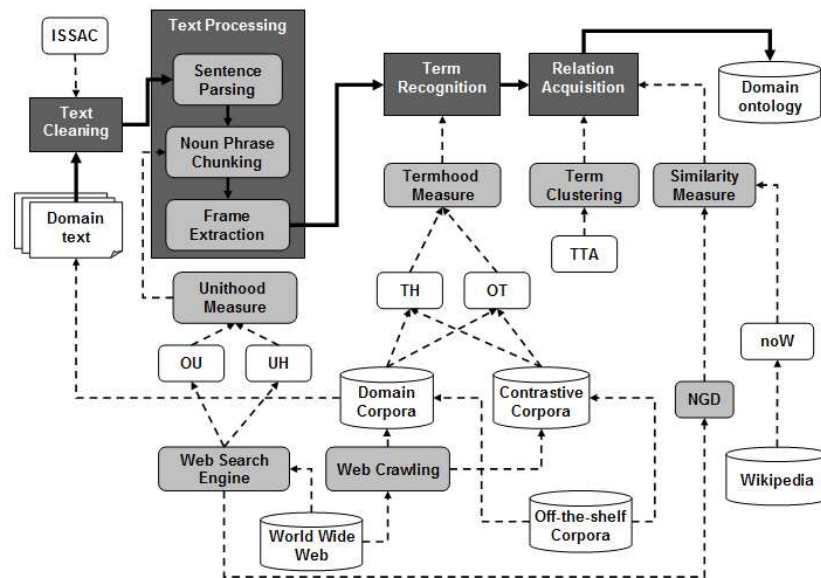


Fig. 1. Ontology construction system architecture. The main phases, namely, *text cleaning*, *text processing*, *term recognition* and *relation acquisition* in the system are represented using shaded rectangles. The rounded rectangles represent the important functionalities required at each phase while the white rounded rectangles depict novel algorithms developed in this research.

corpus. The second type of corpus, known as *domain corpus*, is built through guided web crawling which harvests scientific publications on ScienceDirect using key phrases provided by the domain experts. Electronic versions of domain documents such as textbooks can also make up the domain corpus.

The ontology construction system first performs an optional text cleaning phase using our technique known as *Integrated Scoring for Spelling error correction, Abbreviation Expansion and Case restoration (ISSAC)* [6], [10]. Texts from highly-regarded sources such as academic journals can bypass the text cleaning. Many linguistic analysis tools assume that the input texts are free from spelling errors, abbreviations and improper casings. However, the presence of such noises is inevitable in real-world texts, especially those from online sources. Our text cleaning algorithm known as *ISSAC* has been shown [6], [10] to correct these three types of noise with high accuracy.

Next, domain texts, which can be a selected portion of the domain corpus, are then fed into the sentence parsing module of the text processing phase. The text processing phase is a combination of various natural language processing techniques to extract three-part structures known as *ternary frames*. The modules involved during text processing are *sentence parsing*, *noun phrase chunking* and *frame extraction*. The sentence parsing step unveils part-of-speech tags and dependency structures to enable the identification and grouping of noun phrases during noun phrase chunking. The frame extraction step extracts ternary frames in the form of $\langle arg1, connector, arg2 \rangle$. For the accurate chunking of noun phrases, we incorporated two novel measures known as *Unithood (UH)* [9] and *Odds of Unithood (OU)* [13] for determining the collocation strength of word sequences. Terms are often restricted to noun phrases in many ontology

construction systems. Existing noun phrase chunking techniques typically employ dependency structure analysis and simple word association measures based on static corpora to identify stable sequences of noun phrases. Head nouns having post modifiers or complex sequence of modifiers such as prepositions or conjunctions are often ignored. The inclusion of *UH* and *OU* addresses this problem.

The stable lexical units in *arg1* and *arg2* of the ternary frames generated during text processing are gathered to form a list of term candidates for further processing. The term recognition phase uses both the contrastive and domain corpora to identify terms which are relevant to and representative of the domain of interest. The extent to which the stable lexical units is relevant to some domains is known as *termhood*. The subjective nature of term relatedness or termhood makes term recognition a challenging task to address. Few measures for determining termhood have been developed in the past with limited accuracy. Our two measures, namely, *TH* [8] and *OT* [7] for determining termhood perform with high accuracy in comparison to two other existing measures [12].

During relation acquisition, a flat list of domain terms is organised into a graph structure based on their semantic relations. In this research, we proposed two approaches for acquiring semantic relations between terms. The first approach integrates established and novel techniques in lexical simplification, word disambiguation and association inference for acquiring semantic relations using only dynamic resources on the Web (i.e. Wikipedia and Google) as background knowledge. Our approach queries these resources for information on potential associations, and performs an iterative process of *term mapping* and *term resolution* to identify coarse-grained relations between domain terms. This approach is capable of

handling complex and ambiguous terms, and terms not covered by our background knowledge on the Web. At this stage, more work is required to complete this relation acquisition approach. The second approach utilises cluster analysis to discover unnamed associations between terms. The clustering of terms, unlike other data is less straightforward due to the absence of observable features or attributes. Other linguistic phenomenon such polysemy and synonymy makes term clustering even more challenging. Our *Tree-Traversing Ant (TTA)* algorithm [14] combines the strength of hierarchical clustering with the relatively new ant-based methods to address the problems in term clustering. The use of featureless similarity measures in *TTA*, namely, *noW* [11] and *NGD* [3] enables the accurate estimation of similarity between terms without the need for the computationally expensive task of feature preparation.

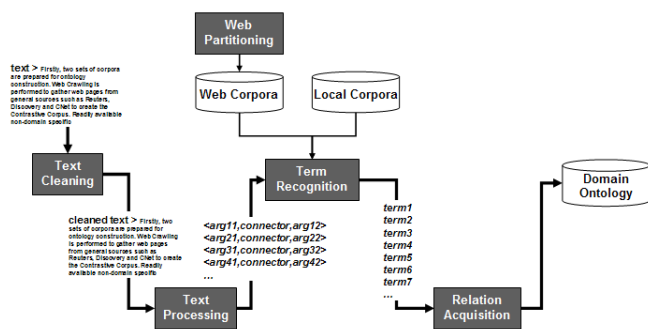


Fig. 2. The intermediate output produced by the system

Figure 2 shows the intermediate output produced after each phase of processing. Our dedicated techniques described above employ only dynamic resources such as Wikipedia and Google. This eliminates the reliance on non-incremental, scarce domain and linguistic knowledge and text corpora during ontology construction.

III. EXPERIMENTS AND DISCUSSIONS

We conducted three independent experiments to assess our ontology construction system in the chemical engineering domain. We prepared three small sets of documents in the following categories: “*chemical compounds*”, “*explosives and toxins*” and “*materials*”. About 8,000 documents were extracted from ScienceDirect.com using keywords provided by experts. These documents together with a process risk management textbook [2] were used as the domain corpus. On the other hand, the contrastive corpus is made up of crawled new articles from Reuters, Discovery and CNet, and off-the-shelf non-domain specific corpora such as GENIA, British National Corpus (BNC) and Reuters-21578. Next, we fed the three sets of domain documents into our system to undergo text processing and term recognition. The outputs after term recognition are 16 terms in the category “*chemical compounds*”, 15 terms in the “*explosives and toxins*” category, and 19 terms from the category “*materials*”. In these experiments, we employ our term clustering algorithm *TTA* to cluster the three

sets of terms for discovering unnamed associations as shown in Figures 3, 4 and 5. At the moment, the associations between the term clusters remain implicit and were left unlabeled. However, we are expecting to address these two issues in our future work on relation acquisition.

We conducted the first experiment by clustering the 16 terms belonging to the category “*chemical compounds*”. This experiment involved both the second and third pass of the clustering approach. Figure 3 shows the results of term clustering from all three passes. After the first pass, the four expected clusters have emerged. In Pass 1 of Figure 3, cluster *A*, *B*, *C* and *D* represents “*photographic chemicals*”, “*disinfectants/biocides*”, “*cosmetic chemicals*” and “*flavour enhancers*”. The prime symbol “*'*” attached to the cluster designators represents the revision or extension of those clusters. For example, cluster *B''* in Pass 3 of Figure 3 is the result of expanding cluster *B'* in Pass 2 of Figure 3 by adding another element “*sodium dichloroisocyanurate*”. While certain chemicals may be used for multiple purposes, the clusters captured by *TTA* represent their dominant applications. While the resulting clusters after the first pass may show some promising results, they have not yet achieved maximum coverage. There are certain qualified terms that have yet to join the appropriate clusters. For example, “*monosodium glutamate*” is a common “*flavour enhancer*”. This is when the second pass contributes to the improvement of the clustering results. As shown in Pass 2 of Figure 3, clusters *B'* and *D'* were appropriately extended to include one additional element each. However, three isolated clusters (i.e. single-element clusters) remain, namely, “*sodium dichloroisocyanurate*”, “*pyrocatechol*” and “*guanosine monophosphate*”. During the final, housekeeping phase, these three isolated clusters were relocated to their intended categories by the third pass as shown in Pass 3 of Figure 3.

The second experiment was conducted using the 15 terms belonging to the category “*explosives and toxins*”. As shown in Figure 4, this experiment was different in that only the first two passes were required to discover the naturally-occurring clusters. We were anticipating three clusters in the final result. Terms from the sub-categories “*nerve agents*”, “*mycotoxins*”, and “*explosives*” were grouped into cluster *A*, *B* and *C*, respectively. An interesting feature revealed in this experiment is *TTA*’s ability to uncover hidden structures within clusters as pointed out by [11]. In Pass 2 of Figure 4, both clusters *C*₁₂ and *C*₃, which are joined by a common ancestor to form cluster *C*^{*P*}, contain the names of the various types of explosives. Within this single parent cluster of “*explosives*”, *TTA* has further discovered two sub-clusters where cluster *C*₃ contains a specific type of explosives known as “*Sprengel explosives*” (i.e. “*Oxylquit*” and “*Panclastite*”).

In the last experiment, we performed clustering using the set of 19 terms from the category “*materials*”. There were four clusters produced by *TTA*, and the cluster memberships properly reflect their intended semantic classes. Cluster *A* consists of the names of the various “*thermoplastics*”, while cluster *B* is made up of several “*silicate minerals*”. Clusters

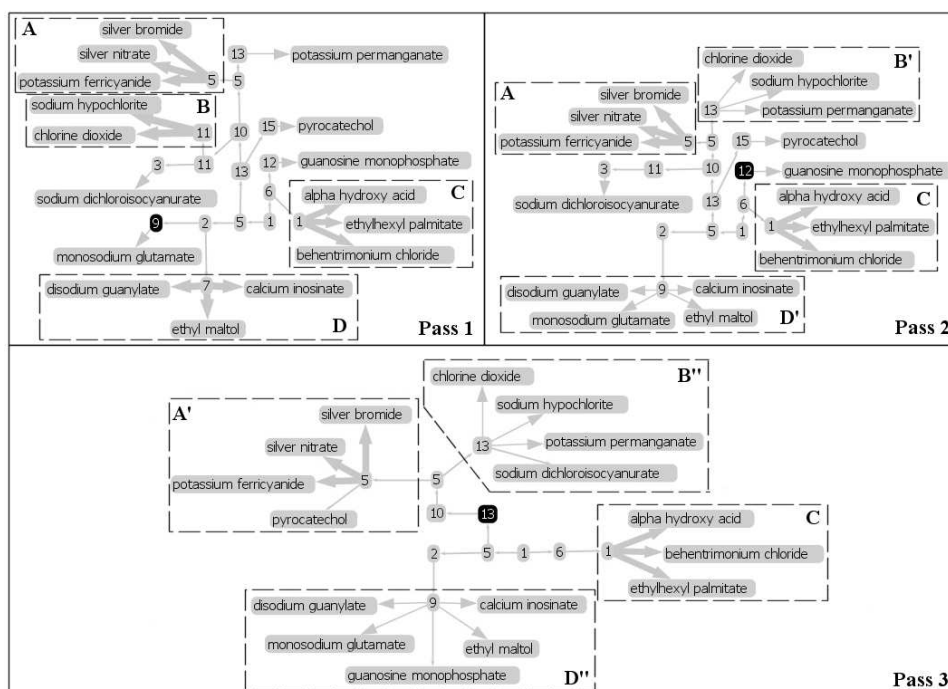


Fig. 3. The result of clustering 16 terms from the category “chemical compounds”. The following thresholds were achieved using the recommended numerical descriptors: $S1_T = 0.71$ and $S2_T = 0.84$. There were four intended clusters of terms from the set. Cluster A, B, C and D, represent “photographic chemicals”, “disinfectants/biocides”, “cosmetic chemicals” and “flavour enhancers”, respectively. A symbol “'” attached to a cluster designator represents a revision of that cluster. A double prime means that the cluster has been revised twice.

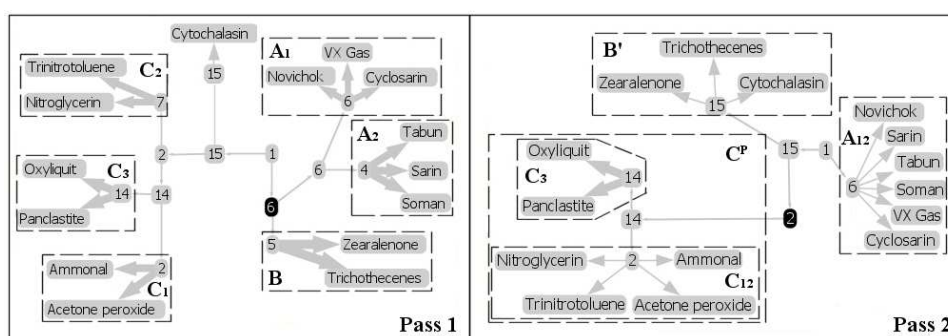


Fig. 4. The result of clustering 15 terms from the category “explosives and toxins”. The following thresholds were achieved using the recommended numerical descriptors: $S1_T = 0.70$ and $S2_T = 0.87$. There were three intended clusters of terms from the set. Cluster A, B and C represent “nerve agents”, “mycotoxins”, and “explosives”, respectively. The symbol “'” attached to a cluster designator represents a revision of that cluster, while the subscript “ij” signifies that cluster X_{ij} is the result of the consolidation of clusters X_i and X_j . The superscript “P” denotes the parent status of the cluster, which is indirectly formed through the sharing of a common node by several clusters belonging to taxonomically-related semantic classes.

C and D contain the names of the several types of “woods” and “glass”, respectively. Similar to the first experiment, the second and third passes were required to finalise the clustering process. Unlike the second experiment where the third pass was of no use, the third pass in this third experiment played an important role. In comparison to the previous two experiments, there were many isolated clusters produced after the first pass. The output after the second pass, as shown in Pass 2 of Figure 5, did not show much improvement. In fact, the intended cluster D did not appear even after the second pass. One of the possible reasons is that the threshold was set much higher than

necessary. As we have mentioned at the start of this section, we utilised the same settings of numerical descriptors to define all thresholds for all experiments. Nevertheless, all remaining isolated terms were properly relocated and the last remaining cluster D, which we had been expecting, had finally emerged as shown in Part 3 of Figure 5.

IV. CONCLUSION

In conclusion, this paper has demonstrated the feasibility and the advantages of our approach to automatically create lightweight ontologies from real-world domain texts. Since our techniques do not rely on static, non-incremental resources

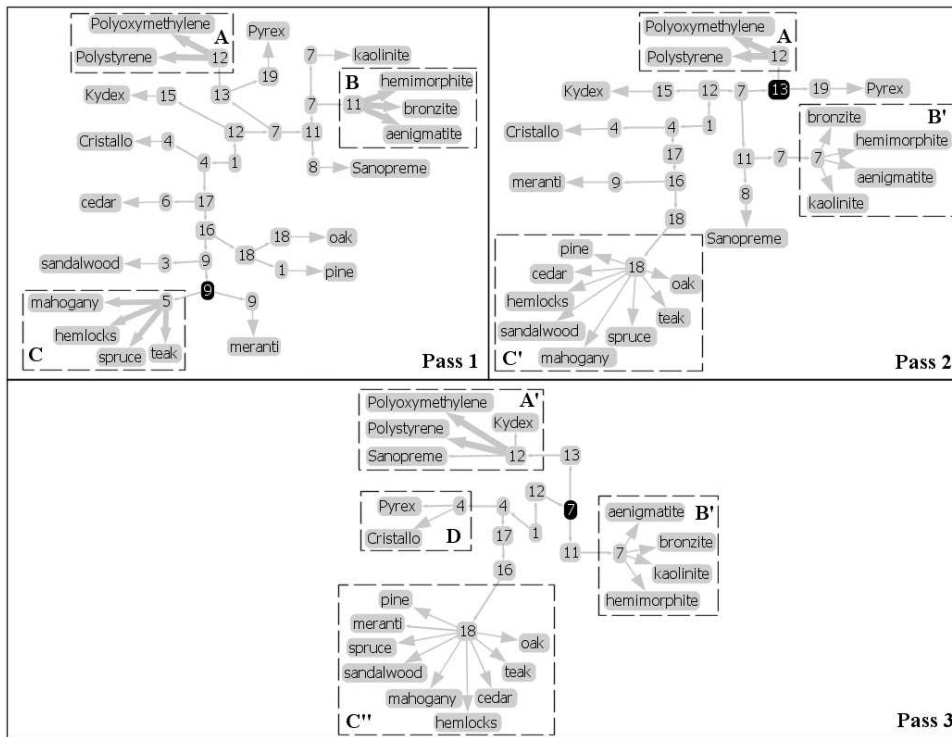


Fig. 5. The result of clustering 19 terms from the category “materials”. The following thresholds were achieved using the recommended numerical descriptors: $S1_T = 0.66$ and $S2_T = 0.85$. There were four intended clusters of terms from the set. Cluster A, B, C and D represent “thermoplastics”, “silicate minerals”, “woods” and “glass”, respectively. The symbol “'” attached to a cluster designator represents a revision of that cluster.

and prescriptions, the system can be extended for use in constructing ontologies for any domains. More work is required to complete the relation acquisition phase to enable the extraction of more complex relations, and to identify the types of semantic relations between terms. We are also planning to evaluate the system using larger data sets.

ACKNOWLEDGEMENT

This research was supported by the Australian Endeavour International Postgraduate Research Scholarship, the DEST (Australia-China) Grant, the 2008 UWA Research Grant, and the Inter-university Grant from the Department of Chemical Engineering, Curtin University of Technology.

REFERENCES

- [1] L. Burnard. Reference guide for the british national corpus. <http://www.natcorp.ox.ac.uk/XMLedition/URG/>, 2007.
- [2] I. Cameron and R. Raman. Process systems risk management. Elsevier Academic Press, 2005.
- [3] R. Cilibrasi and P. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.
- [4] J. Kim, T. Ohta, Y. Teteisi, and J. Tsujii. Genia corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(1):Page 180–182, 2003.
- [5] T. Rose, M. Stevenson, and M. Whitehead. The reuters corpus from yesterday’s news to tomorrow’s language resources. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Canary Islands, 2002.
- [6] W. Wong, W. Liu, and M. Bennamoun. Integrated scoring for spelling error correction, abbreviation expansion and case restoration in dirty text. In *Proceedings of the 5th Australasian Conference on Data Mining (AusDM)*, Sydney, 2006.
- [7] W. Wong, W. Liu, and M. Bennamoun. Determining termhood for learning domain ontologies in a probabilistic framework. In *Proceedings of the 6th Australasian Conference on Data Mining (AusDM)*, Gold Coast, 2007.
- [8] W. Wong, W. Liu, and M. Bennamoun. Determining termhood for learning domain ontologies using domain prevalence and tendency. In *Proceedings of the 6th Australasian Conference on Data Mining (AusDM)*, Gold Coast, 2007.
- [9] W. Wong, W. Liu, and M. Bennamoun. Determining the unithood of word sequences using mutual information and independence measure. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING)*, Melbourne, Australia, 2007.
- [10] W. Wong, W. Liu, and M. Bennamoun. Enhanced integrated scoring for cleaning dirty texts. In *Proceedings of the IJCAI Workshop on Analytics for Noisy Unstructured Text Data*, Hyderabad, India, 2007.
- [11] W. Wong, W. Liu, and M. Bennamoun. Tree-traversing ant algorithm for term clustering based on featureless similarities. *Data Mining and Knowledge Discovery*, 15(3):349–381, 2007.
- [12] W. Wong, W. Liu, and M. Bennamoun. Determination of unithood and termhood for term recognition. In M. Song and Y. Wu, editors, *Handbook of Research on Text and Web Mining Technologies*. IGI Global, 2008.
- [13] W. Wong, W. Liu, and M. Bennamoun. Determining the unithood of word sequences through a probabilistic approach. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP)*, Hyderabad, India, 2008.
- [14] W. Wong, W. Liu, and M. Bennamoun. Featureless data clustering. In M. Song and Y. Wu, editors, *Handbook of Research on Text and Web Mining Technologies*. IGI Global, 2008.