Ant-based Text Clustering Using Semantic Similarity Measure: Progress Report and First Stage Experiment

Haoxiang Xia[1, 2], Shuguang Wang[2], Zhaoguo Xuan[2], and Taketoshi Yoshida[3]
1 Center for Strategic Development of Science and Technology, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
h-xia@jaist.ac.jp
2 Institute of Systems Engineering, Dalian University of Technology, Dalian 116024 China
3 School of Knowledge Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan

ABSTRACT

The ant-based clustering technique has been proven a promising technique for the data clustering problems. However, when applying to text clustering, in many cases the standard ant-based text clustering technique is not satisfactory, partly due to the limitations of the usually adopted VSM-based similarity measure between documents. To address this, a novel ant-based text clustering algorithm is proposed, which utilizes ontology-supported edge-counting-based semantic similarity measure. First stage experiment to test the usefulness and the performance of the proposed algorithm is also reported in the paper.

Keywords: Text Clustering, Ant-based Clustering, Semantic Similarity Measure, Ontology

1. INTRODUCTION

Text Clustering is about partitioning of a set of text documents into self-similar groups so that the elements in the same group or cluster would be more similar than the elements outside the group. This field has become an active research field, in harmony with the increasing demands of categorizing and sorting a large amount of text documents (e.g. in a large database of scientific literature or over the World Wide Web) to assist readers with different interests. In the past decades, various methods have been proposed, such as agglomerative hierarchical clustering [1], k-means [2], Scatter/Gather [3], SuffixTree [4], and Genetic-Algorithms-based approaches [5] as more recent developments. However, due to its intrinsic complexity, text clustering is still an open problem to date.

In the last decade, a novel text clustering approach has been investigated, which is inspired by the collective intelligence-like behaviors of ant colonies. Deneubourg et. al. [6] introduced a clustering algorithm on the basis of the observation of the corpses and larval-sorting activities in ant colonies. In an ant nest, the ants can efficiently heap different items such as corpses and foods in different places to keep the nest neat; it would be then intuitive to consider creating artificial “ants” to make data clusters by imitating the corpses-sorting mechanisms of the real ants. Following such idea, Deneubourg’s general-purposed clustering algorithm was further developed (e.g. [7]) and gradually applied to text clustering (e.g. [8]) as well as to other fields (e.g. graph-partitioning [9]). In parallel with the preceding series of work, another notable ant-behavior-inspired text clustering algorithm is “ANTCLUST”, which was developed by Labroche et al. [10] based on the real ants’ colony-membership recognition mechanism. Their algorithm assigns each item to be clustered to an artificial ant that emits some chemical odor to inditify itself; when those ants randomly meet with each other, they exchange and update their odors on account of the results of the comparison of the similarity between the met ants. Dursing such process, the odors of ants of the same colony become increasingly similar while the odors of ants of the different colonies become more and more dissimilar; and as a result, each ant (i.e. data item to be clustered) would utimately find its best colony.

Such ant-based methods have shown their effectiveness and efficiency in some test cases (e.g., see [11]). However, the ant-based clustering approach is in general immature and leaves big space for improvements. This paper essentially concerns the similarity measurement adopted in the current ant-based text clustering algorithms, which commonly use the cosine measure of vectors in a word vector space, based on the *bag of words* representation of documents. Limitations of such Vector-Space-Model(VSM)-based similarity measure are apparent. On one hand, because of the inevitable diversity of representing words of different documents, the clustering process would inevitably take place in a high-dimenstional space of word vectors. This would tremenously decreases the algorithmic efficiency. To be worse, clustering in a high-dimentional space is difficult because each item tends to have the same distance with all the other items [12]. On the other hand, the semantic meaning of each representing word is neglected in

VSM. Two semantically relevant words would often be treated as two totally different axes in the vector space; and consequently, semantically relevant documents that are respectively represented by these two words may be treated as entirely irrelevant. Since similarity or distance measure serves as a fundemantal criterion for data clustering, it is reasonable to consider that the performance of an ant-based text clustering algorithm would be improved if using a more accurate similarity measure between documents comparing with the currently adopted VSM-based measure.

With this consideration, this paper attempts to explore more accurate similarity measure to improve ant-based text clustering. For this purpose, we consider combining ontology-supported semantic measure with ant-based text clustering, and this paper reports the progress and the primitive results at the current phase of our ongoing efforts in this direction. Section 2 discusses the basic concepts and algorithm of the ant-based text clustering technique. Then, in Section 3, a metric for measuring semantic similarty, which adopts the edge-counting method originated by the work of Rada et.al. [13], is proposed to enhance the ant-based clustering algorithm. The subsequent section gives experimental results of a primitive test case for the proposed method; and the paper ends up with conclusions and discussions of the future work.

2. STANDARD ANT-BASED TEXT CLUSTERING TECHNIQUE

In this section we analyse the general ideas of the ant-based text clustering technique originated by Deneubourg et. al.[6]. In their work, a model was developed to mimic the "clustering" behavior for the *Messor sancta* ants to clean the nests by piling different sorts of items (corpuses, larva, and foods) in different positions. A simple mechanism guides the ants to complete this task: when an ant encounters an item, it tends to pick the item up if the item is dissimilar with the surrounding items; later, if the same ant moves to another position that contains a variety of items that is of the same type of the item being carried by the ant (e.g. the ant carries a dead body to a place that holds a good number of dead bodies), the ant would probably drop the item to that position. With such mechanism, as all the ants in a nest repeat such activities for some period of time, it can be expected that some clusters may be formed with each cluster being comprised of the same type of items.

In Deneubourg et. al.'s model, the prior ant-colony behavior is imitated to perform data clustering. In general, an ant-based clustering algorithm based on Deneubourg et. al.'s model can be described as follows. It first assumes the data objects or items to be clustered are randomly laid down on a two-dimensional $m \times m$ grid or clustering workspace, where m depends on the number of items. Each cell in the grid can contain at most one item. A few artificial "ants" are also placed in the same grid at random. At this initial stage, each ant does not "carry" any item. After completing such initialization process, a cyclic process is designed in which each ant sequentially conducts the following three activities at each step:

1) *Picking up*: At current step, if the ant does not carry any item (i.e. the ant is an "unladen" ant), and if it "encounters" an item o_i (i.e. the ant and the item are located in the same cell at the current step), the ant decides to pick up or ignore that item according to a "picking up" probability P_p, which is a function of local density that determines the similarity of the item o_i with its neighboring items. Less similar items are present, more probably the ant picks the item up.
2) *Moving*: After making the "picking up" decision, the ant randomly moves from the current cell to another cell in the grid. In some variations of the Deneubourg-style ant clustering methods, the ant can only move to an adjacent cell that is not occupied by another ant; but in some other variations, the ant can move across any distance to any other unoccupied cell in the grid.
3) *Dropping*: When the ant reaches a new cell, and if it carries some item (i.e. it is a "laden" ant), the ant requires making another decision whether or not dropping the laden item to this cell, in case that this arrived position does not occupied by another item. Again, the ant calculates another probability (called dropping probability, P_d), which is another function of the similarity between the laden item with the items neighboring this newly-arrived cell. More similar items exist in a local area around the cell, more likely the ant drops the item.

Repeating such activities, the ant may gradually split different types of items into different clusters. The overall process ends when the clusters become stable or the maximal running iteration is reached.

Obviously, the key factors of the above ant clustering algorithm are the picking up and dropping probability functions P_p and P_d. In Deneubourg et. al.'s model, these two functions are determined by defined as the following equations:

Picking up probability, $P_p = (\frac{k_1}{k_1 + f})^2$ (1)

Dropping probability, $P_d = (\frac{f}{k_2 + f})^2$ (2)

$f = f(o_i)$ is a similarity or relevance measure of the item o_i in its neighborhood, while k_1 and k_2 are threshold constants (picking-up threshold and dropping threshold, respectively). P_p is high when $f \rightarrow 0$, indicating that the item o_i would be picked up with a high probability if o_i is dissimilar with its surrounding items. Oppositely, P_d is high when the value if f is high, indicating o_i would probably be dropped to a cell where there are quite some items similar with this item o_i nearby.

For different application contexts, researchers developed different settings for the function $f = f(o_i)$. In the case of text clustering, as the items to be clustered are documents, it is natural to calculate f in accordance with some form of similarity measure between documents. A common document similarity measure is the cosine measure based on the *bag of words* representation of documents. In this similarity measure, a document is represented by a collection of words or index terms with the corresponding weights. For a series of documents, all the representing words form a vector space, and each document can be represented as a vector within this space:

$$d_i = (w_1, w_2, \ldots w_n) \quad (3)$$

Where, n is the dimension of the vector space, and w_1, w_2, …, w_n are the weights of the index terms. w_k=0 if the k-th index term in the vector space is not used to represent the specific document. Consequently, the cosine measure of the similarity between two documents is [14]:

$$sim(d_i, d_j) = \frac{d_i \bullet d_j}{|d_i| \times |d_j|}$$

$$= \sum_{t=1}^{n} w_{i,t} \times w_{j,t} \Big/ \left(\sqrt{\sum_{t=1}^{n} w_{i,t}^2} \times \sqrt{\sum_{t=1}^{n} w_{j,t}^2}\right) \quad (4)$$

And, following [8] (as it provides a typical ant-based text clustering algorithm), the similarity measure between a specific document (denoted by o_i, following the prior description) and the neighboring documents in the grid of the aforementioned ant-based clustering algorithm is:

$$f(o_i) = \frac{\sum_{o_j} sim(o_i, o_j)}{N(o_i)} \quad (5)$$

Where N is the number of cells neighboring the cell where the document o_i is going to be picked up from or dropped to (of course, the definition of neighborhood may be somewhat different in different variations of the ant-based clustering technique). Equation (5) indicates an average similarity between the document o_i and all its neighboring documents.

The previous description portrays the basic ant-based text clustering model of today, and we would like to call it as a “standard” or “typical” ant-based text clustering model. Although currently there are different forms of ant-based text clustering algorithms, in general they share the basic features as described here. On one hand, the clustering process follows Deneubourg’s model; one the other hand, VSM-based representation of documents are applied for similarity measure. As we argued in the introduction section, such VSM-based similarity measure has limitations that would have severe impact on the effectiveness and efficiency of the ant-based text clustering technique. To overcome this and to enhance the performance of ant-based text clustering, we would like to pursue a more accurate similarity measure, seeking help from recent investigations of semantic similarity measure. In the next section, we will describe our attempts to combine semantic similarity measure with the ant-based clustering technique.

3. IMPROVING ANT CLUSTERING WITH SEMANTIC SIMILARITY MEASURE

As we argued previously, the VSM-based similarity measure of documents is a weak point for the standard ant-based text clustering technique. Hence in this section we concern about using more accurate semantic measure to improve ant-based text clustering.

3. 1. Semantic Similarity Measure between Concepts

In the last decade, a number of methods are proposed to measure the semantic similarity between words. In general, these methods can be categorized into two groups, namely edge-counting-based methods and information-content-based methods.

One cornerstone of the edge-counting-based methods is Rada et. al.’s work [13], which proved that the minimal number of edges separating two concepts within a lexical taxonomy of “is-a” links could serve as a metric for measuring the conceptual distance of these two concepts. Following this idea, a various edge-counting similarity measure methods have been proposed. In contrast, following the original work of Resnik [15], the basic idea of the information-content-based methods is to define the similarity of two concepts as the maximum of the information context of the concept that subsumes them in the taxonomy hierarchy. As discussed by various researchers (e.g. [16]), both categories of methods have their advantages and limitations, and this field is still under rapid development.

In this paper, our aim is applying some appropriate semantic similarity measure to enhance ant-based text clustering. To us, as the computational load of the information-content-based similarity measure seems too high, we at the current stage choose the edge-counting approach for semantic similarity measure. Adapted from the method used in [16], the similarity measure we proposed is described as follows.

The foundation of proposed semantic similarity measure is a lexical hierarchy or ontology that is comprised of concepts interconnect with hyponymy ("is-a") links. We take into account two factors for calculating the similarity between two concepts in the ontology: the path length between the two concepts, and the depth of the common ancestor concept (or the "subsumer") in the hierarchy. That means if the two concepts the similarity of which we are going to calculate are c_1 and c_2, then the similarity is a function denoted by Equation (6):

$$sim(c_1, c_2) = f_1(l) \bullet f_2(h) \qquad (6)$$

Where l is the shortest path between c_1 and c_2; and h is the depth of the subsumer of c_1 and c_2 in the ontology. Here in Equation (6) it is assumed that the impacts of parameters l and h on the similarity are independent from one another so that the similarity function is made of two independent functions of f_1 and f_2.

Apparently, the similarity between two concepts would be decreased as the path in the ontology to connect them becomes longer. Furthermore, it would be reasonable to expect the similarity would decrease at an exponential rate; therefore f_1 is defined as Equation (7):

$$f_1(l) = e^{-\alpha l} \qquad (7)$$

where α is real constant that is set between 0 and 1.

The depth of the subsumer is derived by counting the shortest length of links from the subsumer to the top of the ontology (i.e. the root concept). The intuitive observation of the impact of the depth to the similarity measure is that the concepts at higher levels in the ontological hierarchy contain less semantic information content, and therefore two adjacent concepts at a higher level should have less semantic similarity comparing with two adjacent concepts at the lower level. For example, consider a fraction of an ontology about animal classification. Within this ontology, "canine" and "feline" both belong to a "carnivore"; and two lower level concepts of "leopard" and "tiger" both belong to the animal category of "cat", which further belongs to "feline". In this case, people would usually consider that the similarity between "tiger" and "leopard" are higher than that between "canine" and "feline". On account of this observation, a monotonically increasing function of depth h to similarity can then be defined:

$$f_2(h) = \frac{e^{\beta h} - e^{\beta h}}{e^{\beta h} + e^{\beta h}} \qquad (8)$$

where $\beta > 0$ is a constant.

3. 2. Revised Ant-based Text Clustering Algorithm

Based on the prior semantic similarity measure between concepts, this sub-section describes our revised algorithm for ant-based text clustering.

The overall procedure of the proposed algorithm can be described by Fig.1.

procedure ant-based-text-clustering
  Initialize clustering-space and document items
  Initialize ants
  Calculate and store document similarities
  for i=0 *to* Maximal-cycles *do*
    if i/10000 equals an-integer *do*
      adjust-picking-up-threshold
    end-if
    foreach ant do
      decide-pickup
      move
      decide-drop
    end-foreach
  end-for
end-procedure

Fig.1. Overall Algorithmic Procedure

As shown in Fig.1, the proposed algorithm basically follows the same procedure of a standard ant-based text clustering algorithm except that: 1) the similarity between documents is calculated at the initialization stage of the algorithm, instead of in the cycle of ant activities; 2) an "adjust-picking-up-threshold" subroutine is added, aiming at promoting the convergence of the algorithm.

First, let's talk about calculation of the similarity between documents. As depicted in the previous sub-section, calculation of the semantic similarity between two concepts in a lexical hierarchy would probably consume enormous computational resources as it basically involves finding two shortest paths in a network in order to calculate the length between the two concepts and the depth of their subsumer. As a document is represented by a collection of words (concepts), calculation of the similarity between two documents, furthermore, requires comparing multiple pairs of words. For example, if each document is represented by 5 words, the similarity calculation between two documents then requires calculating the similarities of 5*5=25 pairs of words. To be worse, text clustering usually involves a good number of documents, the similarity calculation of pairs of those documents would cause great computational load. It would greatly decrease the algorithmic efficiency if conducting the similarity calculation within the ant activity cycle, since in such case the similarities between documents have to be repeatedly calculated. Our solution is to calculate the similarities ahead of the ant activity cycle, and to store the results to a similarity matrix of pairs of documents. The ants then directly make use of the calculated results in the activity cycle.

One more issue on the similarity measure between documents is about the limitation of the proposed semantic similarity measure. The semantic similarity measure described in the previous sub-section is based on taxonomy. However, in real situations, the developed taxonomy or ontology may not cover all the words (or concepts, in this paper we do not distinguish "words" and "concepts") used to represent documents; and this would decrease the accuracy of the proposed similarity measure. Our strategy is to use a mixed approach that combines the proposed semantic measure with the VSM-based cosine measure.

Assuming that each document is represented by *n* words, we define the semantic similarity of two documents as the average semantic similarity of every pairs of representing words between these two documents:

$$sim_s(d_1,d_2) = \sum_{i=1}^{n}\sum_{j=1}^{n} sim(c_{1,i},c_{2,j}) \Big/ n^2 \qquad (9)$$

Where $sim(c_{1,i},c_{2,j})$ is the semantic similarity measure between two concepts as defined by Equation (6).

The overall similarity between two documents is then a weighted summation of the semantic similarity and VSM-based similarity:

$$sim(d_1,d_2) = \lambda sim_w(d_1,d_2) + (1-\lambda) sim_s(d_1,d_2) \qquad (10)$$

Where $sim_w(d_1,d_2)$ is the cosine measure of similarity between documents d_1 and d_2 as defined by Equation (4); and $\lambda \in [0,1]$ is the weight coefficient.

The second revision of the algorithm is to add the "adjust-picking-up-threshold" procedure or subroutine. Our experiments show that in quite some situations the clustering solution generated by the standard ant-based clustering technique is not very stable, especially when setting a relatively-high picking-up-threshold. When the picking-up-threshold is inappropriately high, the ants may pick items up from well-established clusters and therefore destructed the clusters. As a result, the clusters may cyclically be constructed and destructed, instead of converging into a stable solution. To overcome this, an intuitive tactic is to decrease the picking-up threshold gradually when it is estimated that "good" clusters have formed. This is what the "adjust-picking-up-threshold" subroutine does. Currently, our setting is to decrease the picking-up threshold (k_1 in Equation (1)) to 90.0% of the original level every 10,000 steps, unless the threshold has reached a minimal value 0.001. That is:

procedure adjust-picking-up-threshold
    if k_1>0.001 *do*
        $k_1 \leftarrow 0.90*k_1$
    end-if
end-procedure

Fig.2. Procedure of Adjust-picking-up-threshold

The remaining parts of the algorithm are basically same as the standard ant-based text clustering algorithm. In each step, the ants move randomly within the grid, deciding whether or not to pick up a document, and to drop it down somewhere else, with respect to the probabilities given in Equations (1) and (2). The difference is to replace the similarity measure in Equation (4) with that given in Equation (10).

4. TEST EXEPERIMENT

At present, our work on ant-based text clustering is still in its early stage. Although we have provided an algorithm as described in the previous section, the algorithm's performance has not been sufficiently tested in real world applications or on well-known benchmark data sets. In this section we just present a test result of the proposed algorithm with a relatively small example.

The test experiment is on a collection of Chinese academic articles in the field of business administration, which are sourced from some governmental bureau on planning and coordinating scientific development in

China. 389 documents are randomly selected to test the proposed clustering algorithm. From each document, 5 words (concepts) are extracted so as to represent that document, with the extraction algorithm being presented in a separate paper [17].

Then, to calculate the semantic similarity between documents, we build a shallow ontology to describe the academic branches of business administration, which contains about 1000 concepts interconnected through "is-a" links. As an example, Fig.3 shows a small fragment of this ontology (the concepts are originally in Chinese, being translated to English for the sake of presentation in this paper):

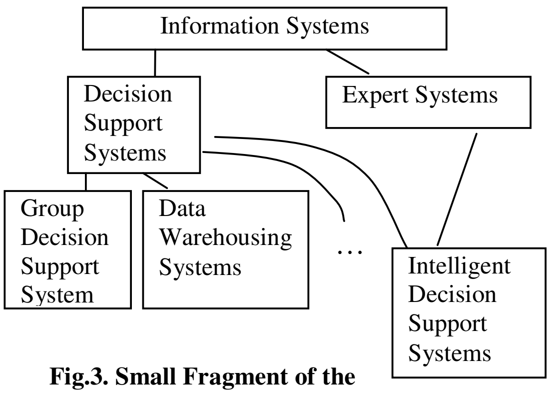

Fig.3. Small Fragment of the Proposed Ontology

We have to acknowledge that the proposed ontology is still a very shallow ontology, not so accurate to describe the destination academic field. Especially, we basically use "is-a" links to describe the relationships between concepts, and it is doubtlessly insufficient and inadequate. In the next stage of work, we will refine the ontology by defining more relationships between concepts, and correspondingly, the semantic measure will also include more semantic relationships.

Making use of this ontology, we test the proposed ant-based clustering algorithm with the selected 389 documents, which are further represented by 1630 Chinese phrases. About 900 of those phrases are described in the proposed ontology. Considering those phrases not included in the ontology, we set λ in Equation (10) to 0.4 so that the cosine similarity measure is also taken into account to some degree. For semantic similarity measure, in this experiment, α in Equation (7) is set to 0.2, and β in Equation (8) is set to 0.6. The picking up threshold (k_1 in Equation (1)) is set to 0.2; and the dropping threshold (k_2 in Equation (2)) is set to 0.05. With these settings, the initial distribution of documents and the final distribution after 30,000 ant steps are respectively shown in Fig.4. and Fig.5.

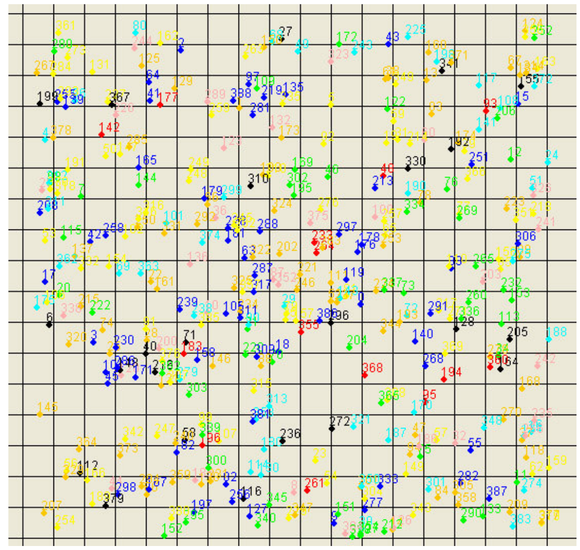

Fig.4. Initial Distribution of Documents In the Test Experiment

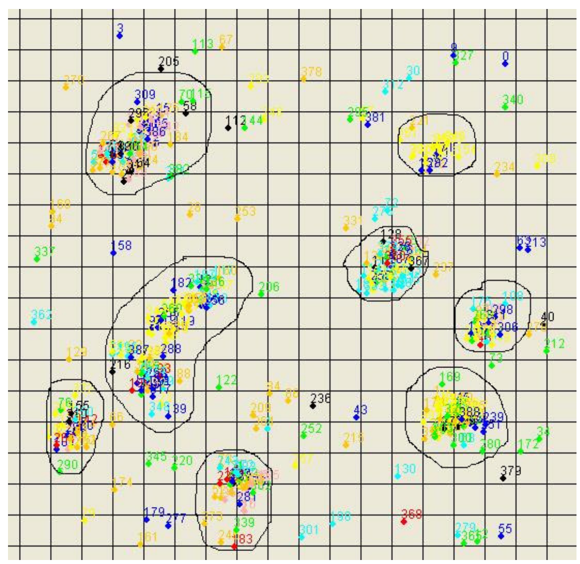

Fig.5. Distribution of Documents After 30,000 Ant Steps

From Fig.5, it can be identified that 8 clusters have formed. The clustering results basically fit the human classification of those documents. In this sense, we can say the proposed clustering algorithm essentially succeeds in partitioning this document set. However, at the current stage, we have not confirmed the contribution of the suggested semantic measure to the overall algorithm. Comparative studies should be conducted in the next stage to further test the usefulness of the proposed algorithm.

5. CONCLUSION

The motivation for us to study the revised ant-based text clustering algorithm proposed in this paper is from our experiments of using the standard ant-based clustering algorithm to partition documents in actual application fields. Our observation is that in many cases the standard ant-based clustering algorithm hardly converges to meaningful clusters. Further investigations show that a principal reason is that the similarities between documents is hard to be reasonably measured by using the VSM-based cosine measure, as different documents seldom share common representing words. With this observation, it is natural to estimate that the performance of the algorithm would be increased by improving the similarity measure with more semantic meanings. Such consideration results in the proposed algorithm in this work.

However, although some experiments (including the experiment described in this paper) have shown that our algorithm looks promising to improve ant-based text clustering, we have to admit that our current experiments have not fully proved the performance of the proposed algorithm. Much work is still required for testing the algorithm. Especially, a major bottleneck of our work is the availability of an ontology with suitable coverage to test the algorithm. At the next stage, we consider making use of the well-established ontology of WordNet [18] to test our algorithm for clustering documents in English.

Furthermore, there are a rich set of different forms of semantic measures in the literature. At the current stage, we are not sure which form is most suited for enhancing ant-based text clustering. The principal reason we adopted an edge-counting semantic similarity measure as described in this paper is its relatively-lower computational load. Further research should be done to investigate more forms of semantic similarity measure.

ACKNOWLEDGEMENTS

This work is partly supported by National Natural Science Foundation of China under Grant 70301009, as well as by Ministry of Education, Culture, Sports, Science and Technology of Japan under "Kanazawa Region, Ishikawa High-Tech Sensing Cluster of Knowledge-Based Cluster Creation Project".

REFERENCES

[1] Ward, J.H. "Hierarchical Grouping to Optimize an Objective Fucntion", Journal of the American Statistical Assoication, 58, p236-244, 1963.

[2] Hartigan, J. and Wong, M. "Algorithm AS136: A k-means clustering algorithm", Applied Statistics, 28, p100-108, 1979.

[3] Cutting D.R., Karger D.R., Pedersen J.O., and Tukey J.W. "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collection", *Proc. ACM SIGIR 92*, p318-329, 1992.

[4] Zamir O. and Etzioni O. "Web Document Clustering: A Feasibility Demonstration", *Proc. ACM SIGIR 98*, p46-54, 1998.

[5] Chiou, Y-C., Lan, L.W. "Genetic Clustering Algorithms", European Journal of Operational Research, 135, p413-427, 2000.

[6] Deneubourg J.L., Goss S., Franks, N. Sendova-Franks A., Detrain C., and Chétien L. "The Dynamics of Collective Sorting: Robot-like Ants and Ant-like Robots", In Proceedings of the 1st International Conference on Simulation of Adaptive Behaviour: From Animals to Animats, 1, p356-363. MIT Press, Cambridge, MA, USA, 1991.

[7] Lumer E. and Faieta B. "Diversity and Adaption in Populations of Clustering Ants", 3rd International Conference on Simulation of Adaptive Behaviour: From Animals to Animats 3. MIT Press, 1994.

[8] Hoe K., Lai W. and Tai T. "Homogeneous Ants for Web Document Similarity Modeling and Categorization", Proceedings of the Seventh International Conference on Parallel Problem Solving from Nature, LNCS 2439. Springer-Verlag, Berlin, Germany, 2002.

[9] Kuntz P. Snyers D. and Layzell P. "A Stochastic Heuristic for Visualizing Graph Clusters in a Bi-dimensional Space Prior to Partitioning", Journal of Heuristics, 5(3):327-351, 1998.

[10] Labroche N., Monmarché N., and Venturini G. "A New Clustering Algorithm Based on the Chemical Recognition Systems of Ants", in Proceedings of 15 European Conference on Artificial Intelligence (ECAI 2002), p345-349, 2002.

[11] Handl J., Knowles J., and Dorigo, M. "On the Performance of Ant-based Clustering", In Design and Application of Hybrid Intelligent Systems, Vol. 104 of Frontiers in Artificial Intelligence and Applications, p204-213, 2003. Amsterdam, the Netherlands: IOS Press.

[12] Beyer K., Goldstein J., Ramakrishnan R., and Shaft U. "When Is 'Nearest Neighbor' Meaningful?" in Proceeding of 7th International Conference of Database Theory (ICDT99), p217-235, LNCS 1540,

Springer-Verlag, Berlin, Germany, 1999.
[13] Rada R., Mili H., Bicknell E., and Bletiner M. “Development and Application of a Metric on Semantic Nets”, IEEE Transactions on Systems, Man, and Cybernetics, 19(1), p17-30, 1989.
[14] Baeza-Yates R., and Ribeiro-Neto B. Modern Information Retrieval, p27-28. Addison-Wesley, 1999.
[15] Resnik, P. “Semantic Similarity in a Taxonomy: An Information-based Measure and Its Application to Problems of Ambiguity in Natural Language”, Journal of Artificial Intelligence Research, 11, 59-130,1999.
[16] Li Y., Bandar Z., and McLean D. “An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources”, IEEE Transactions on Knowledge and Data Engineering, 13(4), p871-882, 2003.
[17] Xuan Z., Dang Y., Jiang X. and Zhao M., “A High Precision Algorithm for Automatic Extracting of High-frequency Words Based on Statistics”, In this Proceedings, 2005.
[18] Miller G.., “WordNet: a lexical database for English”, Communications of the ACM, 38 (11), p39-41, 1995.