

Clustering Annotations of a Digital Text with Bio-inspired Approaches

¹Maiga Chang, ²Rita Kuo, ³Keith Ying, ⁴Andrew F. Chiarella, ⁵Jia-Sheng Heh, ⁶Kinshuk

^{*1}*Athabasca University, maigac@athabascau.ca*

²*Knowledge Square, Ltd., rita.mcs1@gmail.com*

³*Chung-Yuan Christian University, keith70167@hotmail.com*

⁴*Athabasca University, andrewc@athabascau.ca*

⁵*Chung-Yuan Christian University, jsheh@ice.cycu.edu.tw*

⁶*Athabasca University, kinshuk@athabascau.ca*

Abstract

Students often annotate texts while reading by highlighting, underlining, adding marks, and writing comments beside the text. These may serve various functions and will reflect each student's goals and understanding of the text. The research team has proposed two simple biology-inspired approaches to represent the patterns of student annotations and to help hierarchical clustering method classify students based on the similarity between their annotations; the annotations produced were simple highlighting. Both approaches have instant processing speed (i.e., less than 100 ms for dealing with a batch of forty students' annotations on a 2894-word text) compared with generic hierarchical clustering algorithm implemented in Matlab and reasonable clustering accuracy (i.e., both the precision and recall values are in the range of 0.67 to 0.71) compared with the clusters created by human raters. This research further enhances the bio-inspired annotation clustering methodology by proposing two new approaches, the Cosine and the Diffusion approaches, to solve the problems earlier approaches have as well as to improve clustering accuracy.

Keywords: *Biology-Inspired; Chromosome; Patterns; Clustering*

1. Introduction

Annotation is useful for students to enhance their learning from text [2]. In one common form of annotation, students read text and highlight the words they believe important in order to make their learning more efficient. Due to the popularity of electronic books and tablet PCs (e.g., Android tablets and iPad), many researchers have developed web-based annotation tools [12][25] or reading software that supports such annotation.

Annotations made by different students are usually not same and can be considered as a behavior pattern which represents each student. For this reason, the research team has proposed two bio-inspired approaches, the Standard and the Quantitative approaches, to convert the highlights made by a student into a bit-string chromosome and to measure the similarity of two chromosomes [26]. The annotation data for highlights of a text can be represented as a data sequence of 0's and 1's. The two existing approaches transform the students' highlights on the text to bit-string chromosomes, then locate and quantify chromosome differences between students and cluster students with similar chromosomes.

If students highlight the text in similar ways then they may be clustered into the same group. In other words, the students in the same cluster may have similar thoughts on the text or similar learning strategies. The clustering results may then be used to inform students of annotations they had not considered that had been completed by other students similar to themselves.

In order to assess the performance of the proposed chromosome-based approaches, the research team first compared the processing time used by each approach with that used by the generic hierarchical clustering method (implemented in Matlab). Furthermore, the accuracy rate of each approach was checked by comparing the results with the clusters identified by human raters. Although the results were good in both processing speed and cluster accuracy, the Standard and the Quantitative approaches have some shortcomings. First, in the Standard approach each difference is also weighted by position. Therefore, the total difference value for two pairs of chromosomes could be very large (i.e., the exponentially increasing difference value issue); also, the difference value for two pairs of chromosomes with the *same* number of differences can be quite different (e.g., differences later in the

sequence are given more weight than others) and a different set of differences could produce the same overall difference score (i.e., the indistinguishableness issue). Second, although the Quantitative approach can solve the exponentially increasing difference value issue the Standard approach has, it suffers from the indistinguishableness issue.

This research proposes two new approaches, the Cosine and the Diffusion approach, to solve the problems the other two approaches have in representing students' highlights on the text and measuring the difference for chromosomes. With the new difference measurement, the agglomerative hierarchical clustering method may also have more accurate results when use human rater's clusters as benchmark. The remainder of the paper is organized as follows: Section 2 reviews the literature on annotation and clustering approaches. Section 3 discusses the two enhanced approaches. Section 4 describes the method of clustering chromosomes. Section 5 shows the evaluation results. Section 6 provides a conclusion and discusses possible future research directions.

2. Annotation and Clustering

Annotation can help students read and learn. According to Marshall (1997), there are telegraphic, or implicit, and explicit annotations [19]. The implicit annotations (e.g., highlighting and underlining) contain meanings that generally only the original annotator could know. The explicit annotations (e.g., note taking and drawings) have more clearly stated meanings that other readers may also understand. As electronic books and tablet PCs have become more popular, digital and electronic documents are widely used and accepted by students. Making annotations on electronic documents for learning has gradually become more attractive to students [2]. There are many tools for making annotation on electronic and on-line documents today: *Annotate* is a commercial software application which allows multiple users to do annotations on images and documents online collaboratively [1]; Chao and colleagues (2010) developed a cross-media annotation system [5]; Chiarella and colleagues designed a reading application, called *CoREAD*, that supports simple highlighting annotations and, in addition, adds social text signals to the text by aggregating the annotations of the readers [6][7]; Kurhila and colleagues (2003) developed an annotation system called *EDUCOSM* for web-based learning [16]; *comment* is a commercial web-based annotation tool which allows multiple users to annotate text and to write text-based comments online [8]; Rau and colleagues (2004) developed Web-based Annotation Tools (WATs) for students making annotations directly on the web pages and sharing their annotations with others [22]; and, Yeh and Lo (2009) developed *Online Annotator* for EFL writing which allows teachers to mark errors and to do corrections on online documents and students can receive the corrective feedback from teachers in an online composition class [25].

Some research has been conducted to assess the effectiveness such tools have for improving students' learning performance. Chang and colleagues (2006) have found that annotating learning materials is an effective learning strategy for student's reading comprehension ability [4]; Hwang and colleagues (2007) have found that the student's learning performance does improve in both individual learning and group learning scenarios [13]; and, Hwang and Hsu (2011) have found that pre-reading and annotation sharing positively effect students' learning [12].

This research aims to represent students' annotation behaviors using a chromosome-like string and then classify these chromosomes into different clusters. Clustering is a means used to divide a data set into subsets. The items in a subset are similar or have the same features as the other items [9][24]. Many clustering methods exist. Two well-known clustering methods are Hierarchical Clustering and Partitioning Clustering [10][17]; we employ agglomerative hierarchical clustering with Ward's method to cluster the data according to the difference values for chromosomes measured by the two approaches.

In Hierarchical Clustering a hierarchical tree is constructed that presents the clusters in a tree structure called a dendrogram [3][15]. In the hierarchical tree, the root is a cluster which contains all items from the data set and the leaves are the individual items. Figure 1(a) shows a data set with six items and Figure 1(b) shows a possible hierarchical tree for the data set. Different hierarchical trees can be established with different similarity measurement approaches.

3.1. Cosine Approach

One issue that both the Standard and the Quantitative approaches have is that the value representing the difference between two chromosomes may become very large. To overcome this limitation, the research team treats each student's chromosome as a vector and uses cosine similarity to calculate the difference between two vectors. Cosine similarity works by calculating the value for the cosine of the angle between the two vectors. The cosine may take values between -1 and +1; where +1 indicates perfect similarity. Take two angles in Figure 5 as example, when User 1 and User 2's behaviors are transformed to the vector form, the two vectors may compose an angle θ_1 . Similarly, User 1 and User 4's annotations in vector form compose angle θ_2 . The value for the cosine of angle θ_1 is closer to +1 than angle θ_2 , which means, User 2's behavior is more similar to User 1's.

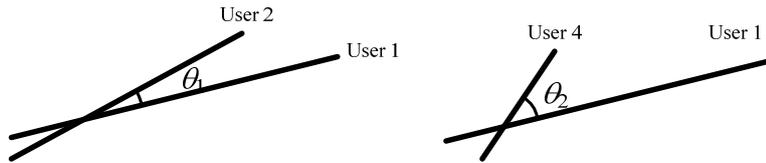


Figure 5. Angles composed by two vectors

Although the Cosine approach does not produce exponentially increasing values for the differences when comparing two users, the approach does have a similar problem to that of the Quantitative approach (i.e., the indistinguishableness issue). In Figure 6, the differences between User 1 and User 2 and User 1 and User 3 are shown. After we apply the Cosine approach we find the cosine similarity between User 1 and User 2 is 0.8571 and the cosine similarity between User 1 and User 3 is also 0.8571. The results suggest that the annotations made by Users 2 and 3 are identical to each other. However, we can see that User 2 and 3 are quite different when we compare them directly, as Figure 7 shows. The reason that both of User 2 and User 3 have the same cosine similarity value is they both have the same number of differences when compared with User 1. In order to overcome this limitation the research team designed and developed the Diffusion approach described next.

User #1	01100101111011011110001	User #1	01100101111011011110001
User #2	00100111110111011110001	User #3	01100101111110111110000

(a) Chromosomes represent User #1 and User #2 (b) Chromosomes represent User #1 and User #3

Figure 6. Chromosome comparisons

User #2	00100111110111011110001
User #3	01100101111110111110000

Figure 7. Comparing User #2 and User #3

3.2. Diffusion Approach

The Diffusion approach is inspired by diffusion phenomena such as when a drop of watercolor paint lands on a piece of tissue paper as shown in Figure 8(a). The "hot zone" is the original area where the paint landed. After a while, the water will spread, or diffuse, to the surrounding area as indicated by the "diffusion" area in Figure 8(a). In the Diffusion approach, however, to simplify the process of identifying diffusion area of a hot zone, we only read the diffusion phenomena from one direction (i.e., the right hand side of the hot zone) as Figure 8(b) shows.

4. Clustering Algorithm

For student clustering, this research applies agglomerative hierarchical clustering to cluster the chromosomes by mapping the chromosomes on a two-dimensional plane. This research divides a chromosome into two parts and uses each part's difference value calculated by the proposed bio-inspired approaches as the x-y coordinates in order to put the chromosome on a two dimensional plane, as Figure 11 shows.

	x-axis	y-axis
User #1	011001011110	11011110001
User #2	001001111101	11011110001
User #3	011001011110	11011110001
User #4	111001011110	10111110000

Figure 11. Axis values of Users #1 to #4's chromosomes

Take User 1 as the benchmark chromosome for the coordinates calculation. Assuming User 1's coordinates is (1, 1), then we can have User 2's coordinates is (0.7143, 1) according to the cosine similarity values calculated with the Standard approach [26]; User 3's coordinates is (1, 1); and, User 4's coordinates is (0.9354, 0.7715). At the very beginning, each user represents a single user cluster at very beginning. After that, a new cluster C_5 can be found by combining cluster C_1 and C_3 together. The coordinates of cluster C_5 's center is still (1, 1) due to both C_1 and C_3 locate at same place. We can the merge cluster C_5 with C_4 and have a larger cluster C_6 with center coordinates (0.9677, 0.8858). At the end, the cluster C_2 and C_6 are merged and have a root cluster C_7 with center coordinates (0.841, 0.9429). Figure 12 shows the hierarchical tree based on the similarity measured by the Cosine approach.

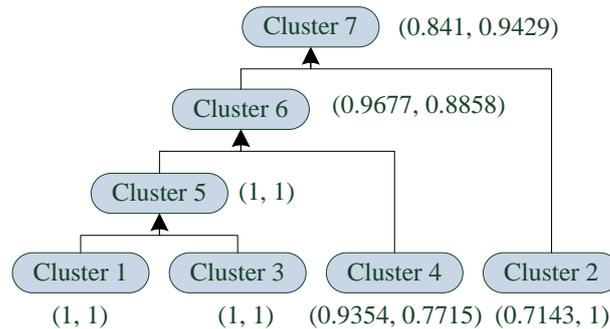


Figure 12. Hierarchical tree based on the similarity measured by the Cosine approach

The Diffusion approach uses a different way to cluster chromosomes. The approach uses diffusion areas and the sum of 1's in the areas as coordinates. Take Figure 13 as example, there are five diffusion areas in the User 1's chromosome: the first diffusion area is "01", therefore, the x-axis of this diffusion area is 1 and the y-axis of this diffusion area is the sum of 1's in the area (i.e., only one 1 exists). However, because the second diffusion area is "0" and there is no "1" in it, no coordinate assigned for it. Based on this rule, the coordinates of the followed diffusion areas are (3, 1), (4, 1), and (5, 1). At the end, the four coordinate values are sum up and the coordinate of the entire chromosome becomes

$$(1+3+4+5, 1+1+1+1) = (13, 4)$$

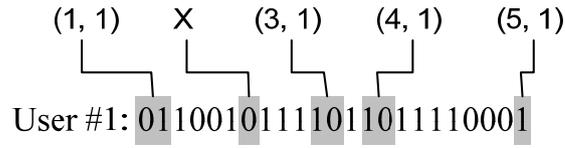


Figure 13. Temporary coordinates of the diffusion areas

Applying the abovementioned rule, we can have User 2's coordinates as (14, 4); User 3's coordinates as (13, 4); and, User 4's coordinates as (8, 4). With the bottom up clustering mechanism, Figure 15 shows the hierarchical tree based on the similarity measured by the Diffusion approach.

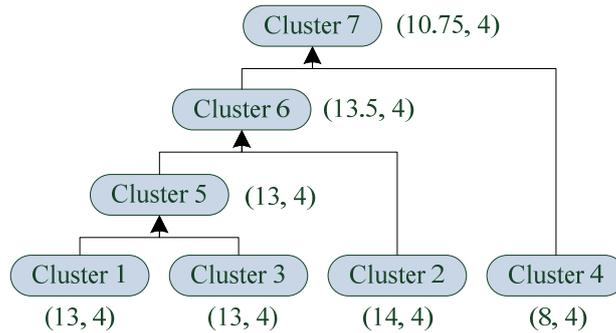


Figure 15. Hierarchical tree based on the similarity measured by the Diffusion approach

5. Evaluation

To evaluate the two enhanced chromosome-based clustering approaches two performance indicators were used: speed (i.e., how fast the approach can process the data) and performance (i.e., how well the task is performed). The research team implemented a generic hierarchical clustering method with pdist, linkage, and dendrogram functions in Matlab and compared the processing time across the different approaches when used to cluster the 40 students' annotations on the 2894-word text. Table 1 lists the processing times that each approach took.

Table 1. Time Cost

	<i>Hierarchical Clustering with Cosine Similarity</i>	<i>Hierarchical Clustering with Diffusion Similarity</i>	<i>Generic Hierarchical Clustering</i>	<i>Human Rater</i>
<i>mapping (ms)</i>	46.18	<u>0.78</u>	N/A*	N/A*
<i>clustering (ms)</i>	132.52	<u>19.75</u>	28.43	263,690
<i>Total (ms)</i>	178.7	<u>20.53</u>	28.43	263,690

In this research, only one human rater is involved and asked for clustering students manually according to their annotations of the text. The human rater prints the annotated text out and sorts it into different clusters with his own judgment. Similar to the hierarchical clustering, the research team also asks him to further combine two or more clusters together to form a larger one. At the end, we have the clusters classified by human rater as the benchmark.

The Cosine approach is the slowest approach when compared with others besides the human rater because the vector computation and floating point multiplication take a longer time. The Diffusion approach, on the other hand, is the fastest approach even compared with the two earlier approaches (i.e., the Standard approach's 29.95ms and the Quantitative approach's 27.43ms) [26]. As well, the Diffusion approach takes less time than the generic hierarchical clustering approach.

For the performance evaluation, the research team considered the results clustered by the human rater as the benchmark or ideal outcome. Although it is arguable that this research only has one human rater, it is understandable that it is difficult to find high agreement among human raters and multiple human raters may make the comparison complicated and less meaning. The results (i.e., the clusters) produced by the three approaches (i.e., the Cosine approach, the Diffusion approach and the generic hierarchical clustering Matlab approach) are therefore compared with the human rated clusters to determine which performed the best.

Two widely used measures for evaluating the performance of classification tasks from the information retrieval research field are precision and recall. The modified definition of the precision measure for this research is

$$precision = \frac{|manual_cluster_users| \cap |approach_cluster_users|}{|approach_cluster_users|} \quad (1)$$

Precision means the accuracy of the clustering results that an approach produces. For example, if there are ten users in the proposed cluster classified by approach A and only three users exist in the same cluster classified by the human cluster, then the precision of approach A is 0.3 (i.e., only 30% of the clustering results are correct).

Precision may not be enough for evaluating an approach's performance because precision only takes the output of an approach into consideration. For example, when all users in a proposed cluster classified by approach A occur in the cluster classified by the human cluster, the precision of approach A is 1. However, what if we know that there are only three users in the cluster created by approach A but there are twenty users in the cluster classified by the human rater? Approach A no longer appears to be a highly performing technique because it missed 17 of the users as classified by the human rater.

The recall provides a means to assess any misses that occurred in the abovementioned example. The modified definition of the recall measure for this research is

$$recall = \frac{|manual_cluster_users| \cap |approach_cluster_users|}{|manual_cluster_users|} \quad (2)$$

Take the abovementioned case into consideration, only this time we will use the recall measure to evaluate the performance of approach A. Since there are twenty users in the cluster classified by the human rater and all three users in the proposed cluster classified by approach A are correct, the recall of approach A is only 0.15 according to Eq.(2). This is significantly less than the perfect score of 1 that approach A gets for its precision.

Once again, the recall measure is not perfect and works independently. Considering approach A has fifty users in a cluster and the cluster classified by the human rater only has twenty users; if all of the latter users are in the cluster classified by approach A, then approach A gets 1 for its recall. But once again, approach A is not a good approach because 30 users classified in the cluster are not correct classifications according to the human rater.

It is easy to use precision and recall, but we should not consider only one at a time; that is why we need the F-measure [11][18]. The F-measure is a weighted combination of precision and recall and has a range of 0 to 1.

$$F_{\beta} = (1 + \beta^2) \times \frac{precision \times recall}{\beta^2 \times precision + recall} \quad (3)$$

The F-measure has been used to evaluate the performance of clustering methods [14][21][23]. The beta value can be adjusted to favor more precision or more recall. The beta value should be set close to 0.5 (i.e., $F_{0.5}$) to favor precision and close to 2 (i.e., F_2) to favor recall. The research team also used the F-measure to evaluate the performance of the two enhanced approaches. Table 2 lists the precision, recall, and F-measure values for each approach.

Table 2. Average * precision, recall and F-measure for each clustering approach

	<i>Hierarchical Clustering with Cosine Similarity</i>	<i>Hierarchical Clustering with Diffusion Similarity</i>	<i>Generic Hierarchical Clustering</i>
<i>Precision</i>	<u>0.7488</u>	0.7047	0.6843
<i>Recall</i>	0.6877	<u>0.7162</u>	0.6668
<i>F₂</i>	0.6814	<u>0.6979</u>	0.6513
<i>F_{0.5}</i>	<u>0.7062</u>	0.6903	0.6590

*. Beside direct use of 2894-bit string chromosomes, we also shorten the chromosome length by considering sentence-based and paragraph-based annotation. We consider a sentence is annotated if the amount of highlighted words in the sentence exceeds a certain threshold (percentage), e.g., 10%. Therefore, we cluster the data set a hundred times with different sentence and paragraph threshold combinations and the numbers in Table 2 are the average results.

The generic hierarchical clustering approach has minimum values of precision and recall compared with the proposed two approaches. The two approaches have better performances than the earlier two in different ways: the Cosine approach has the highest precision than others (i.e., the Standard approach's 0.7146 and the Quantitative approach's 0.7027) and the Diffusion approach has the highest recall compared to the others (i.e., the Standard approach's 0.7061 and the Quantitative approach's 0.6747) [26].

6. Conclusions

From the results, we see that the Cosine approach has highest precision among all approaches; however, it also takes a lot of time. On the other hand, the Diffusion approach has the highest recall and is also the fastest approach. Since the Diffusion approach's $F_{0.5}$ value is quite reasonable if compared with the Cosine approach but almost nine times faster than the Confusion approach, we may consider the use of the Diffusion approach for clustering students' annotations.

According to Table 1, the Diffusion approach is capable of clustering 40 students' annotations on 2894-word text in 21 milli-seconds. Such processing speed makes the Diffusion approach feasible to be implemented as real-time web and cloud service. For instance, a quick-response recommender system can be developed by using the Diffusion approach to automatically provide a student with the shared thoughts, bookmarks, key discussions, internal/external references for further readings, and important notes made by classmates or even past students whose annotations are classified to the same cluster.

Such system can be implemented as a plug-in for learning management systems (e.g., Moodle). They may be implemented in annotation plug-ins, such as Marginalia for Moodle. Moreover, such systems could work with e-reader applications (e.g., Kindle, Apple iBooks app, and Kobo) to provide learners with a valuable service for self-learning and reading activities.

In this research, the data set size is relatively small (i.e., 40 student annotations on a 2894-word text). We do not know if the processing time for mapping and clustering will increase linearly or exponentially as the data set size grows. It is important to figure out the relationships among the chromosome length, the data set size, and the processing time in the future.

Acknowledgment

The authors wish to thank the support of Athabasca University, the Mission Critical Research funding, NSERC, iCORE, Xerox and the research related gift funding provided to the Learning Communities Project by Mr. Allan Markin.

References

- [1] Annotate documents and images online, Retrieved August 15, 2012, from <http://a.nnotate.com/>
- [2] E. Ball, H. Franks, J. Jenkins, M. McGrath, J. Leigh, "Annotation is a valuable tool to enhance learning and assessment in student essays", *Nurse Education Today*, vol. 29, no. 3, pp. 284-291, 2009.

- [3] J. Bien, R. Tibshirani, "Hierarchical Clustering with Prototypes via Minimax Linkage", *Journal of the American Statistical Association*, vol. 106, no. 495, pp. 1075-1084, 2011.
- [4] C.-K. Chang, G.-D. Chen, C.-K. Chen, "Using computer-based annotation to promote the summary of reading comprehension", *Communication of IICM*, vol. 9, no. 1, pp. 41-58, 2006.
- [5] P.-Y. Chao, G.-D. Chen, C.-W. Chang, "Developing a Cross-media System to Facilitate Question-Driven Digital Annotations on Paper Textbooks", *Journal of Educational Technology & Society*, vol. 13, no. 4, pp. 38-49, 2010.
- [6] A. F. Chiarella, "Enabling the collective to assist the individual: A self-organising systems approach to social software and the creation of collaborative text signals", Unpublished PhD Thesis, Montreal, McGill University, 2008.
- [7] A. F. Chiarella, S. P. Lajoie, "A self-organising systems approach to history-enriched digital objects", In D. Ifenthaler, P. Pirnay-Dummer & N. M. Seel (Eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge*, pp. 131-158, 2010, Boston, MA: Springer.
- [8] Co-ment - Text annotation and collaborative writing, Retrieved August 15, 2012, from <http://www.co-ment.com/>
- [9] X. Cui, J. Gao, T. E. Potok, "A flocking based algorithm for document clustering analysis", *Journal of Systems Architecture*, vol. 52, no. 8-9, pp. 505-515, 2006.
- [10] C. Ding, T. Li, "Adaptive dimension reduction using discriminant analysis and K-means clustering", In *Proceedings of the 24th international conference on Machine learning*, #521, 2007. Retrieved on August 13, 2012, from <http://www.machinelearning.org/proceedings/icml2007/papers/521.pdf>
- [11] G. Hripcsak, A. S. Rothschild, "Agreement, the F-Measure, and Reliability in Information Retrieval", *Journal of the American Medical Informatics Association*, vol. 12, no. 3, pp. 296-298, 2005.
- [12] W.-Y. Hwang, G.-L. Hsu, "The effects of pre-reading and sharing mechanism with annotations on learning", *The Turkish Online Journal of Educational Technology*, vol. 10, no. 2, pp. 234-249, 2011.
- [13] W.-Y. Hwang, C.-Y. Wang, M. Sharples, "A study of multimedia annotation of Web-based materials", *Computers & Education*, vol. 48, no. 4, pp. 680-699, 2007.
- [14] A. K. Jain, "Data clustering: 50 years beyond K-means", *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651-666, 2010.
- [15] L. Khan, M. Awad, B. Thuraisingham, "A new intrusion detection system using support vector machines and hierarchical clustering", *The International Journal on Very Large Data Bases*, vol. 16, no. 4, pp. 507-521, 2007.
- [16] J. Kurhila, M. Miettinen, P. Nokelainen, P. Floréen, H. Tirri, "Peer-to-Peer Learning with Open-Ended Writable Web", In *Proceedings of the 8th Annual Conference on Innovation and Technology in Computer Science Education*, pp. 173-177, 2003.
- [17] M.-J. Li, M.-K. Ng, Y.-M. Cheung, L.-Z. Huang, "Agglomerative Fuzzy K-Means Clustering Algorithm with Selection of Number of Clusters", *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 11, pp. 1519-1534, 2008.
- [18] D. D. Lewis, W. A. Gale, "A sequential algorithm for training text classifiers", In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp.3-12, 1994.
- [19] C. C. Marshall, "Annotation: from paper books to the digital library", In *Proceedings of the second ACM international conference on Digital Libraries*, pp. 131-140, 1997.
- [20] U. Neisser, "Rising scores on intelligence tests", *American Scientist*, vol. 85, no. 5, pp. 440-447, 1997. Retrieved on August 13, 2012, from <http://www.americanscientist.org/issues/pub/rising-scores-on-intelligence-tests/1>
- [21] M. Rasmussen, G. Karypis, "gcluto: An interactive clustering, visualization, and analysis system", Technical Report CSE/UMN TR 04-021, Department of Computer Science and Engineering, University of Minnesota, 2004.
- [22] P.-L. P. Rau, S.-H. Chen, Y.-T. Chin, "Developing web annotation tools for learners and instructors", *Interacting with Computers*, vol. 16, no. 2, pp. 163-181, 2004.
- [23] M. Steinbach, G. Karypis, V. Kumar, "A Comparison of Document Clustering Techniques", In *Proceedings of KDD-2000 Workshop on Text Mining*, 2000.
- [24] R. Xu, D. Wunsch, II, "Survey of clustering algorithms", *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645-678, 2005.
- [25] S.-W. Yeh, J.-J. Lo, "Using online annotations to support error correction and corrective feedback", *Computer & Education*, vol. 52, no. 4, pp. 882-892, 2009.
- [26] K. Ying, M. Chang, A. F. Chiarella, Kinshuk, J.-S. Heh, "Clustering Students based on Their Annotations of a Digital Text", In *Proceedings of the 4th IEEE International Conference on Technology for Education*, pp. 20-25, 2012.