Visualization and Clustering of Document Collections using a Flock-based Swarm Intelligence Technique

Richard H. Fowler, Raul A. Huerta, and Wendy A. L. Fowler
Department of Computer Science, University of Texas – Pan American, Edinburg, TX, USA

Abstract *– Electronic availability of documents continues to increase, yet identifying documents relevant to the user remains a primary constraint in electronic document use. Visual representations of document collections can facilitate search by representing large collections of documents in a manner that is complementary to linear, text based representations. Visual representations can provide a means to make the overall structure of a collection comprehensible, as well as a mechanism to identify groups of useful documents and access relevant individual documents. The current work employs flock-based clustering to both organize documents and provide visual representations of documents. Reynolds' three rule flocking scheme is augmented with additional rules to provide document clustering. A unified visual representation supplies facilities for overview of the entire document collection, filtering documents, and retrieving individual documents. The system also utilizes visualization tools for individual cluster identification and exploration based on keyword search. Stereoscopic viewing is provided to enhance users' perception of 3D organization.*

Keywords: visualization, information visualization, information search, clustering, swarm intelligence

1 Introduction

Electronic availability of documents continues to increase both with respect to types of access and breadth of coverage. Yet, identifying those documents that are relevant to the user for the task at hand remains the primary constraint in electronic document use. Most systems provide access using keyword search, whether using publicly available Internet search engines or digital libraries accessed by professionals. Visual representations of document collections during search can augment text-based techniques by representing large collections of documents in a manner that is complementary to linear, text based representations. Visual representations can provide a means to make the overall collection comprehensible, as well as mechanisms to identify groups of useful documents and access relevant individual documents.

In information visualization tasks the essential approach is to provide "overview first, zoom and filter, then details-on-demand" [24]. The approach to visualization of document collections we have taken follows that paradigm closely. The current work describes a system that provides a visual representation of a complete document collection formed using a biologically inspired flock-based [21] clustering technique [4, 5]. The technique provides a means to not only form clusters of documents, but also spatially order the clusters and documents within clusters. Together, these spatial orderings can provide the user both a global view of the document collection, as well as the ability to view relations at a more detailed intra-cluster level. Simultaneous cluster formation and spatial arrangement is efficient by eliminating the need for separate computational stages.

The system we have developed employs flock-based clustering to both organize documents with respect to content and to provide visual representations of documents. The basic clustering approach whereby additional rules are added to Reynolds' three rule flocking scheme [21] was augmented and tuned for the web based document sets with which the work was done. A single visual representation provides users facilities to gain an overview of the entire document collection, filter the document collection, and obtain information about individual documents. The system also provides visualization tools for individual cluster identification and exploration based on keyword search. Both desktop and large screen stereoscopic viewing facilities are provided to enhance the users' perception of three dimensional organization.

The following sections first present Related Work in swarm intelligence used to form clusters, focusing on flock-based techniques. This section also describes the metrics utilized to represent individual documents and document collections and extensions to Reynolds' model that have been employed with document collections. The next section describes the System that we have developed for flock-based visualization of document collections. Finally, Conclusions are presented and References listed.

2 Related Work

Clustering is the process of assigning a set of objects into groups, or clusters, so that the objects in the same cluster are more similar to each other than to those in other clusters [11]. For document retrieval the clustering of documents in a collection is widely used to facilitate user directed search through browsing [14]. Clustering techniques that make use of flock-based swarm intelligence techniques are relatively new and provide a useful adjunct to other techniques, such as hierarchical and nearest neighbor clustering.

2.1 Flock-based Modeling

Reynolds' seminal work in creating flock-like visual behaviors of birds [21] was designed to provide a technique for use in computer animation. As such, its goal was to provide a visual representation of flight and the grouping of individuals, i.e., flocking, that would appear realistic to viewers, rather than provide a biologically accurate model of behavior. Reynolds' agent-based technique was quite successful using a very limited number of rules. The basic flocking approach was quickly extended to a wide range of groups, including herds [7], schools of fish [9], crowds [29], unmanned air vehicles [30], and robots [31]. Further refinements have added additional rules to provide for flock-like behavior in the presence of predators, obstacles to avoid, and path following [22]. The approach has also been extended to multiple species flocking [5, 17] using techniques closely related to those for document clustering. The next sections present details of the original flocking model from which later extensions are derived, followed by a discussion of extensions to the basic model with a focus on clustering and document collections.

2.1.1 Simulating Flock Motion

Reynolds' original technique to create perceptually realistic flocking behavior for animation was based on providing rules for movement, or steering, for individual agents. Only three rules are used, illustrated in Figure 1, which are applied by each individual of the flock to steer its movement (direction and velocity):

- Alignment: Steer towards the average direction of movement of nearby agents
- Separation: Steer to avoid being too close to nearby agents
- Cohesion: Steer to move toward the center position of nearby agents.

Using these three rules, individual agents initially placed at any spatial location will adjust their directions of movement and velocities to form a single group that exhibits movement perceptually quite similar to a flock of birds or other social biological group, e.g., school of fish. These steering rules are applied considering only other agents within a small range around each individual, as shown by the circles in the figure, rather than to all other agents. A velocity vector is formed for each of the three rules, alignment, separation, and cohesion. These vectors are then used to determine the individual agent's velocity.

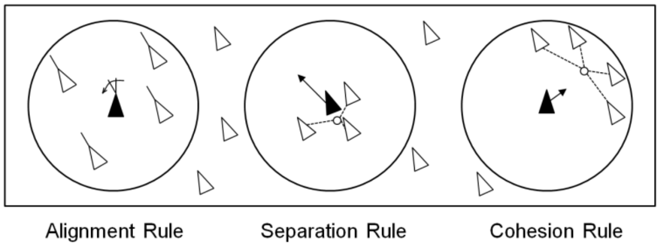

Figure 1. Only three basic rules of movement are necessary to create perceptually realistic flocking motion. The filled triangle represents the agent applying the rules, outline triangles show other agents, and the circle represents the maximum spatial range of rule evaluation.

The alignment rule orients the direction of movement for an agent with the direction of nearby agents. Nearby agents are considered those that are within the range of evaluation, $d(P_x, P_a) \leq d_1 \cap d(P_x, P_a) \geq d_2$, where d is distance, P is position, x denotes another agent, and a the agent for which alignment is being determined. d_1 and d_2 determine the other agents that are considered in determining alignment and are pre-defined. d_1 represents the maximum range of rule evaluation, and d_2 is used to exclude the nearest other agents, which will be affected by the separation rule. Each agent's velocity vector, $\vec{v}_{align}$, is aligned with the average of the velocity vectors $\vec{v}_x$ for the n agents within the range of evaluation:

$$\vec{v}_{align} = \frac{1}{n}\sum_{x=1}^{n} \vec{v}_x \tag{1}$$

The separation rule keeps an agent from colliding with another agent. It does this by changing the agent's velocity vector, $\vec{v}_{sep}$, depending on distance from agents in the range of evaluation, $d(P_x, P_a) \leq d_2$:

$$\vec{v}_{sep} = \sum_{x=1}^{n} \frac{\overline{\vec{v}_x + \vec{v}_a}}{d(P_x, P_a)} \tag{2}$$

The cohesion rule orients the movement of an agent toward the center of nearby agents. The agent's velocity vector, $\vec{v}_{coh}$, is oriented to the direction of the average spatial position of the agents within its evaluation range of evaluation, $d(P_x, P_a) \leq d_1 \cap d(P_x, P_a) \geq d_2$:

$$\vec{v}_{coh} = \sum_{x=1}^{n} \overrightarrow{(P_x - P_a)} \tag{3}$$

Finally, the velocity vector for each agent, $\vec{v}_a$, is calculated by summing and weighting the velocity vectors calculated by the three rules:

$$\vec{v}_a = w_{align} \cdot \vec{v}_{align} + w_{sep} \cdot \vec{v}_{sep} + w_{coh} \cdot \vec{v}_{coh} \tag{4}$$

with w_{align}, w_{sep}, and w_{coh} pre-determined weights.

2.2 Extensions of Flock Based Modeling to Document Collection Clustering

By the agents' iteratively applying the three rules of alignment, separation and cohesion, a single group forms that moves in a perceptually good approximation of group movement of identical biological entities in nature. However, there are other cases of interest that the rules do not capture. For example, in order to create movement of herd animals across terrain, Gompert [7] augments Reynolds' three rules modeling bird flight with a fourth rule whereby each agents' position also determined by the elevation of the terrain at the agent's position. Such derivation of rule sets appropriate to the problem at hand is characteristic of agent based modeling [15].

In addition to single group modeling, another question that has been addressed using flock-based modeling is differentiation of the members of a single group of agents into distinct groups. This would occur, for example, when agents model birds, but the birds are of different species. In this case each species would form its own separate group, and each separate group would exhibit the movement patterns captured by Reynolds' three rules. Solutions to this problem add additional rules whereby agents, e,g., birds, are brought closer together or pushed farther apart depending on their similarity. One way to determine similarity values is by comparing feature vectors representing individual agents. Agents of the same species share many features and so have high similarity, which leads to their spatial positions moving closer. These groups of similar agents are also moved away from other, dissimilar, agents, which have also formed groups based on high inter-agent similarity. This general approach, in which additional rules are added to flock-based clustering, has been used to provide multiple groupings, or clusters, of individual interests [20], time varying data [17], arbitrary attributes [19, 26], and spatial data [6].

Clusters of documents can be identified using the same approach in which additional rules are added that consider similarity among feature vectors [5, 6]. For document clustering, additional rules can be added that consider the similarity of the topic content of documents, where the feature vector is a vector of terms used to describe a document's content. The next section describes techniques for determining document feature vectors and their similarity. In the following section techniques for deriving document clusters based on inter-document similarity are presented.

2.2.1 Document Feature Vectors and Similarities for Clustering

The most widely used measures of similarity among documents are based on the Vector Space Model [23]. This technique utilizes a document's words to transform each document to a feature vector representation that captures the document's content. Comparisons among documents' feature vectors are then employed to provide document similarities.

The Vector Space Model uses a list of indexing terms defined for a particular document collection. These terms can come from a fixed vocabulary or be derived for individual document sets. The text of each document is analyzed, and a vector, *d*, representing each document is created. *d* is of length equal to the number of indexing terms. Each element of *d* is given a weight, *w*, for each indexing term, *i*. Using this method, each document can be considered as a single point in a space of dimensionality equal to the number of index terms.

Typically, inter-document similarity considers the number of terms common to a document pair as represented by their term vectors, a measure of their content similarity, normalized by number of terms in documents. In order to increase the discriminative power of terms, terms in the term vector *T* are first weighted by the inverse frequency of occurrence in the term set, the *tf*×*idf* model [27]. The idea is that relatively rarely occurring terms are more useful in characterizing inter-document similarity, and, so, are differentially weighted. Equation 5 below provides the weight, *w*, for a term in the document term vector. *tf* is a term's frequency in the document collection, and tf_{ik} is the number of occurrences of term T_k in d_i. *N* is the number of documents in the collection and n_k represents the number of documents containing term T_k. The weight, *w*, of term T_k is:

$$w = \frac{tf_{if} \times log\left(\frac{N}{n_k}\right)}{\sqrt{\sum_{j=1}^{T_n}\left(tf_{if}\right)^2 \times \log\left(\frac{N}{n_j}\right)^2}} \qquad (5)$$

2.2.2 Flock-based Document Clustering

The earliest use of flock-based clustering was as one element of a hybrid clustering approach that used conventional techniques together with flock-based techniques for final determination of clusters [6]. At that time other swarm based techniques for clustering were also being explored, including ant colony optimization [13] and particle swarm optimization [16]. The first use of flocking-based techniques for document collections was by Cui et al. [5, 6]. Refinement of the approach continues, with recent work on efficient implementation demonstrating the clustering of 500,000 documents [32]. Contemporary with the work of Cui et al, Picarougne and colleagues developed a flock-based clustering system with visualization facilities [18, 19].

As noted, the basic approach of flock-based document clustering is to augment Reynolds' original three agent rules with additional rules. Cui et al. [5] utilized a document's term vector representation as the feature vector from which values for inter-document similarities were calculated, and these were then used in additional rules.

To determine document clusters Cui et al. use two variations of rules to augment Reynolds' original approach to determining spatial placement of agents. The basic goal is to move similar documents together and dissimilar documents apart by adding two rules. In the first rule, strength of attraction, $\vec{v}_{sim}$, is proportional to the distance between the agents and the similarity between the agents' term values:

$$\vec{v}_{sim} = \sum_{x=1}^{n} \left(S(A,X) \times d(P_x, P_a) \right) \tag{6}$$

where *S(A,X)* is the similarity value determined using the term vectors, or feature sets, for agents A and X. Similarly, the strength of repulsion, $\vec{v}_{dis}$, is inversely proportional to the distance between the agents and the similarity between the agents' features:

$$\vec{v}_{dis} = \sum_{x=1}^{n} \frac{1}{S(A,X) \times d(P_x, P_a)} \tag{7}$$

As before, the velocity vector for each agent, v_a, is calculated by summing the weighted velocity vectors calculated by the, now five, rules:

$$\vec{v}_a = w_{al} \cdot \vec{v}_{al} + w_s \cdot \vec{v}_s + w_c \cdot \vec{v}_c + w_{sim} \cdot \vec{v}_{sim} + w_{dis} \cdot \vec{v}_{dis} \tag{8}$$

with w_{al} , w_s, w_c ,w_{sim}, and w_{dis} pre-determined weights.

This basic approach was refined [6] by combining the two additional rules into a single rule that uses a variable threshold, *T*, by which attractive and repulsive forces can be manipulated based on agent feature similarity:

$$\vec{v}_{sim-threshold} = \sum_{x=1}^{n} \frac{(S(A,X) - T) \times \overrightarrow{(P_x - P_a)}}{d(P_x, P_a)} \tag{9}$$

There is now a single velocity vector based on document similarity, v_{ds}, that depends on the pre-defined threshold, *T*, to determine document similarity based agent movement. Such changes to rules are characteristic of extensions made to the basic flock based algorithm.

3 System

The system we have developed provides a visual representation of document collections based on flock-based clustering as one element of a document retrieval system. To perform the clustering, documents are first represented as term vectors. Then, using flock-based modeling that extends Reynolds' original approach, documents with high similarity move together and dissimilar documents move apart in a three dimensional space. These movements create spatial clusters of documents that share content, as represented by their term vectors, and clusters move apart spatially to fill the space that users view, as shown in Figure 2 for a set of web retrieved documents. In addition to the main viewing window, views from the top and sides of the document space are also available to provide orientation. Figure 3 shows the view after the user has zoomed in on two of the clusters. Further zooming allows the inspection and retrieval of individual documents.

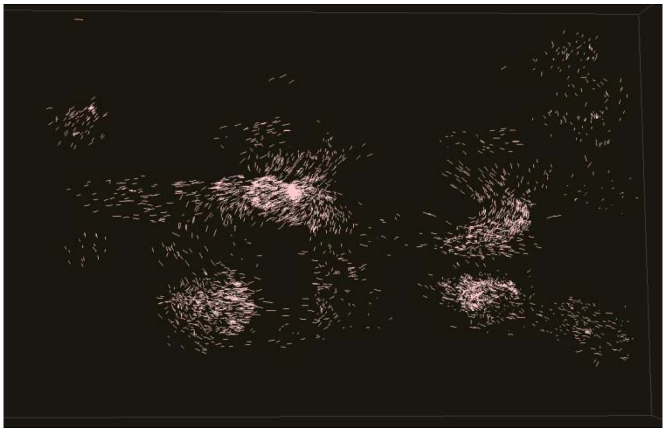

Figure 2. Flock-based clustering for a document collection. User's are supplied tools to browse the collection, identify clusters of interest, and retrieve individual documents.

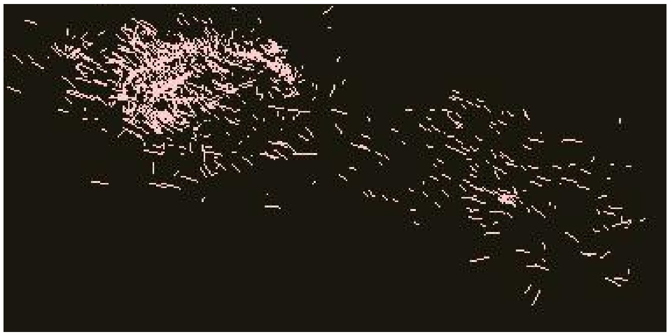

Figure 3. User has zoomed in on the two clusters at the lower right of the display. Further zooming allows selection of individual documents for display.

Stereoscopic viewing is available to enhance perception of the three dimensional space in which documents are displayed. In various task-based studies user's performance has been shown to be enhanced through stereoscopy [1, 8, 28]. Additionally, the user can interactively change the view orientation, e.g., "spinning" and "jittering" the display, to discern further information about the nature of the spatial clustering [25].

As with other flock-based document clustering systems [5, 32], additional rules for agent movement based on document similarities are used by the system to augment Reynolds' three principle steering rules. A variant of Cui et al.'s threshold-based similarity rule [6] is used for clustering, providing a threshold for document separation that can be adjusted to tune cluster formation and separation for different document sets. The flock-based clustering employed by the system represents documents using the *tf*×*idf* method, using the most frequently occurring terms in document sets. Similarity between agents based on feature vectors is calculated as $sim_{ij} = \sum_{k=1}^{T} w_{ik} \times w_{jk}$, where inter-document similarity, sim, for each document pair, i, j, for each of T elements in the term vector, is derived from term weights, w, for agent pairs. For efficiency it is calculated once prior to initiating flocking and stored, rather than recalculated on each iteration.

One tenet of information search is that users should be provided multiple paths of access to information [2]. Though the capabilities of the system center on flock-based clustering and its visual representation, conventional keyword search for individual documents is also available and provides a useful addition to cluster based search [3, 12]. The system's flock-based clustering visualizations provide one element of its document retrieval functions. Other elements are designed to support an iterative document retrieval process in which users' information needs are defined and met [10].

In addition to retrieving individual documents through keyword search, the system also provides facilities that combine keyword search and visual cluster representation. The goal is to help users know where to look and explore within the large visual representation in order to find clusters that contain documents likely to be relevant to the user's information needs. This cluster identification is accomplished by augmenting the cluster display by visually marking individual documents that match a keyword search. Users are then visually directed to those clusters with many keyword matches and which contain similar documents, as indicated by common cluster membership.

Another means by which flock-based cluster visualizations could provide capabilities integrated with a suite of document retrieval mechanisms is by using it to provide an alternative representation of search results provided by a conventional search engine. Internet search engines return results ranked with respect to relevance to a keyword based query. Were the query able to express the user's information need exactly and the retrieval mechanism able to then supply the documents that met the user's information need, then there would be little to desire in such a system. Unfortunately, this is not the case, due in part to users' inability to completely specify information needs in terms used by the retrieval system. Rather, it is more likely that only some degree of the user's need is met with initially retrieved documents. Further refinement of information needs and search vocabulary are parts of the iterative process information retrieval. By using the flock-based visualization system to provide clustered, versus sequential, ordering of documents within a retrieved set, the user could be provided a visual mechanism complementary to the sequence of documents to find relevant documents through exploration of the clusters of documents formed from a set of retrieved documents. Additionally, flock-based clustering is particularly efficient for incremental clustering [32], and the set of visually displayed document can be increased and maintained as the search continues.

4 Conclusions

The current work extends the use of flock-based clustering visualization in document retrieval through its integration with tools supporting the iterative information retrieval process. The system provides mechanisms for cluster identification by keyword query match to identify individual documents and show their location in the complete document collection, thereby enabling users to efficiently explore the document space. This exploration facilitates user information need specification, an important component in the retrieval process, as well as individual document retrieval. The system provides multiple paths to information items through its facilities for document collection browsing, cluster identification, and keyword based search.

5 Acknowledgements

This work was supported by National Science Foundation grant MSI-CIEC OCI-0636352 and the University of Texas – Pan American Computing and Information Technology Center.

6 References

[1] K. W. Arthur, K. S. Booth, and C. Ware, "Evaluating 3D task performance for fish tank virtual worlds," *ACM Transactions on Information Systems*, vol. 11, no. 3, pp. 239-265, 1993.

[2] M. J. Bates, "Subject access in online catalogs: A design model," *Journal of the American Society for Information Science*, vol. 37, no. 6, pp. 357-386, 1986.

[3] D. B. Crouch, C. J. Crouch, and G. Andreas, "The use of cluster hierarchies in hypertext information retrieval," in *Proceeding of Hypertext '89*, pp. 225- 237, 1989.

[4] X. Cui, J. Gao, and T. E. Potok, "A flocking based algorithm for document clustering analysis," *Journal of Systems Architecture*, vol. 52, no. 8-9, pp. 505-515, 2006.

[5] X. Cui and T. E. Potok, "A distributed agent implementation of multiple species flocking model for document partition clustering," *CIA 2006, Lecture Notes in Computer Science*, vol. 4149, pp. 124-137, 2006.

[6] G. Folino and G. Spezzano, "An adaptive flocking algorithm for spatial clustering," *Parallel Problem Solving in Nature (PPSN) VII, Lecture Notes in Computer Science*, vol. 2439, pp. 924-933, 2002.

[7] J. Gompert, "Real-time simulation of herds moving over terrain," *Proceedings of Artificial Intelligence and Digital Entertainment*, pp. 149-150, 2005.

[8] N. Greffard, F. Picarougne, and P. Kuntz, “Visual community detection: An evaluation of 2D, 3D perspective and 3D stereoscopic displays,” *Proceedings of 19th International Symposium on Graph Drawing, Lecture Notes in Computer Science*, vol. 7034, pp. 215-225, 2011.

[9] Y. Inada, “Steering mechanism of fish schools,” *Complexity International*, vol. 8, pp. 1-8, 2001.

[10] P. Ingwerson and I. Wormwell, “Improved subject access, browsing and scanning mechanisms in modern online ir,” *Proceedings of ACM SIGIR*, pp. 68–76, 1986.

[11] A.K. Jain, M.N. Murty, and P.J. Flynn, “Data clustering: A review,” *ACM Computing Surveys*, vol. 31, no. 3, pp. 264-323, 1999.

[12] N. Jardine and C.J. van Rijsbergen, “The use of hierarchical clustering in information retrieval,” *Information Storage and Retrieval*, vol. 7, pp. 217–240, 1971.

[13] N. Labroche, N. Monmarche´, G. Venturini, “AntClust: Ant clustering and web usage mining,” *Proceedings of Genetic and Evolutionary Computation Conference*, pp. 25–36, 2003.

[14] A. Leuski, “Evaluating document clustering for interactive information retrieval,” *Proceedings of the International Conference on Information Knowledge and Management (CIKM)*, pp. 33-40, 2001.

[15] C. M. Macal and M. J. North, “Tutorial on agent-based modeling and simulation,” *Journal of Simulation*, vol. 4, pp. 151-162, 2010.

[16] V.D. Merwe and A.P. Engelbrecht, “Data clustering using particle swarm optimization,” *Proceedings of IEEE Congress on Evolutionary Computation*, pp. 215–220, 2003.

[17] A.V. Moere, “Information flocking: time-varying data visualization using Boid behaviors,” *Proceedings of the Eighth International Conference on Information Visualization*, pp. 409–414, 2004.

[17a] S. Momen, B. P. Amavasai, and N. H. Siddique, “Mixed Species Flocking for Heterogeneous Robotic Swarms,” *EUROCON 2007 The International Conference on Computer as a Tool*, pp. 2329-2336, 2007.

[18] F. Picarougne, H. Azzag, G. Venturini, and C. Guinot, “On data clustering with a flock of artificial agents,” *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004)*, pp. 777-778, 2004.

[19] F. Picarougne, H. Azzag, G. Venturini, and C. Guinot, “A new approach of data clustering using a flock of agents,” *Evolutionary Computation*, vol. 15, no. 3, pp. 345-367, 2006.

[20] G. Proctor and C. Winter, “Information flocking: Data visualisation in virtual worlds using emergent behaviours,” *Proceedings of Virtual Worlds*, pp. 168–176, 1998.

[21] C. Reynolds, “Flocks, herds, and schools: A distributed behavioral model,” *Computer Graphics*, vol. 21, no. 4, pp. 25-34, 1987.

[22] C. Reynolds, “Steering behaviors for autonomous characters,” *Proceedings of Game Developers Conference*, pp. 763–782, 1999.

[23] G. Salton, C. Yang, and A. Wong, “A vector space model for automatic indexing,” *Communications of the ACM,* vol. 18, no. 11, pp. 613-620, 1975.

[24] B. Shneiderman, “The eyes have it: A task by data type taxonomy for information visualizations,” *Proceedings of IEEE Symposium on Visual Languages*, pp. 336-343, 1996.

[25] B. W. van Shooten, E. M. A. G. van Dijk, E. Zudilova, A. Suinesiaputra, and J. H. C. Reiber, “The effect of stereoscopy and motion cues on 3D interpretation task performance,” *Proceedings of the International Conference on Advanced Visual Interfaces*, pp. 167-170, 2010.

[26] E. Sklar, C Jansen, J. Chan, and M. Byrd, “Toward a methodology for agent-based data mining and visualization,” *International Workshop on Agents and Data Mining Interaction (ADMI 2011)*, pp. 20-31, 2011.

[27] K. Sparck-Jones, “A statistical interpretation of term specificity and its application in retrieval,” *Journal of Documentation*, vol. 28, no. 1, pp. 11–20, 1972.

[29] W. Tang, T. R. Wan, and S. Patel, “Real-time crowd movement on large scale terrains,” *Proceedings of the Theory and Practice of Computer Graphics (TPCG’03)*, pp. 146-153, 2003.

[29] C. Ware and G. Franck, “Evaluating stereo and motion cues for visualizing information nets in three dimensions,” *ACM Transactions on Graphics*, vol. 15, no. 2, pp. 121-140.

[30] N. R. Watson, N. W. John, and W. J. Crowther, “Simulation of unmanned air vehicle flocking,” *Proceedings of Theory and Practice of Computer Graphics*, pp. 130-137, 2003.

[31] N. Xiong, J. He, J. H. Park, T. Kim, and Y. He, “Decentralized flocking algorithms for a swarm of mobile robots: Problem, current research and future directions,” *6th IEEE Consumer Communications and Networking Conference (CCNC 2009)*, pp.1-6, 2009.

[32] Y. Zhang, F. Mueller, X. Cui, and T. Potok, “Data-intensive document clustering on graphics processing unit (GPU) clusters,” *Journal of Parallel and Distributed Computing*, vol. 71, pp. 211-224, 2011.