Adaptive Visual Clustering for Mixed-Initiative Information Structuring

Hakan Duman, Alex Healing, and Robert Ghanea-Hercock

British Telecommunications plc, Adastral Park, Martlesham Heath,
Ipswich, IP5 3RE, United Kingdom
{hakan.duman,alex.healing,robert.ghanea-hercock}@bt.com

Abstract. Cyclone is a mixed-initiative and adaptive clustering and structure generation environment which is capable of learning categorization behavior through user interaction as well as conducting auto-categorization based on the extracted model. The strength of Cyclone resides in its integration of several visualization and interface techniques with data mining and AI learning processes. This paper presents the intuitive visual interface of Cyclone which empowers the user to explore, analyze, exploit and structure unstructured information from various sources generating a personalized taxonomy in real-time and on-the-fly.

Keywords: Information Visualization, Data Mining, Human-Computer Interaction, Machine Learning.

1 Introduction

The structuring of information is a fundamental step in order to deal with the sheer volume now common in the digital universe and increasing at an accelerating pace [1]. The World Wide Web has acted as a catalyst, making it easier for both information providers to publish and consumers to link themselves to a virtually infinite array of heterogeneous sources of information. With such a huge resource at the fingertips of practically anyone with an internet connection, never before has it been more relevant to investigate means to deal with the potential information deluge by employing intelligent approaches to automating the process. Data mining is capable of extracting patterns in large corpuses of data that would be unfathomable for a human to work with, yet automated processes alone are insufficient, as the way in which they work to retrieve "interesting" features for their human users must, by definition, be user-driven. Rather than for the prescription and processing stages to act distinctly, a mixed-initiative approach, integrating them simultaneously, has several advantages in communicating to the human the results of the automated process as well as allowing the user direct manipulation with both the data and the automated process itself [2].

The field of interactive data mining [3],[4],[5] focuses on interfaces which can facilitate this form of bidirectional communication between automated processes and human, primarily consisting of three key goals: (1) *data visualization* - exploiting humans' pattern-matching abilities aiding understanding and to discover "interesting"

M.J. Smith and G. Salvendy (Eds.): Human Interface, Part I, HCII 2009, LNCS 5617, pp. 384–393, 2009.

features; (2) *direct manipulation* - to allow the user to specify more easily what they seek in terms of the automated processes (e.g. selecting a subset of the dataset for the automated data mining processes to analyze at a more fine-grained level); (3) *visibility of both the process and result of the data mining to the user* - minimizing misinterpretation [3],[6]. We have identified a fourth important goal, closing the loop once more between human and machine: *offering a means for the user to feedback a performance measure of the data mining to the system* - thus tailoring the way that the automated processes operate and visualizing this *adaptation* in real-time. In addition to specifying which data for the automated mining to act upon, this fourth element influences *how* the automated processes take place.

There has been a significant amount of work on applying data mining techniques in the context of information structuring, however the focus of this research has often been on the performance of the data mining algorithms and less about the human-computer interface and user experience [7],[8],[9],[10]. Commercial off-the-shelf products such as Verity, Gammaware, Vivisimo and Inxight offer tools for automating taxonomy generation and maintenance, however, the majority of these products are relatively heavyweight in terms of their processing and the data mining and machine learning employed typically remains a "black box" to users. Although [11],[12] in particular provide learning the categorization behavior from the user, the way in which this is accomplished lacks the visibility and flexibility we argue is necessary to ensure users understand and trust the system during this process. The commercial approaches mentioned above all provide a visual interface to aid understanding of the relationships between information as well as allow the user to explore, analyze and exploit it. However, this visualization is limited to the visualization and manipulation of the resulting taxonomy and facilities to increase the users' awareness of the automated data mining and categorization processes are neglected.

We introduce an interactive data mining environment called *Cyclone* which has an emphasis on clustering and structuring the information inputted, generating a taxonomy. The primary method of visualization in Cyclone is similar to that of [3],[13],[14], consisting of a force-based graph of nodes representing information points; the forces and subsequent distances between the nodes reflecting their derived similarity. We evaluate to what extent this approach is an effective means to address the four elements identified above, conveying the result of machine-based data mining employed but also acting as a means to *adapt* the data mining process based on direct manipulation by humans. Using adaptive visual clustering, the interface presents to the user how the automated data mining is based on both unsupervised (statistical) measures of the data in combination with a history of user actions as learnt through their interaction. This is particularly relevant for the information structuring task we focus on, where the categorization of information might be highly user-specific, as well as the interpretation of the data dynamic as the task or information changes. By introducing a supervised learning component into our system, and this being trained by the direct manipulation of data by the user in an intuitive manner, our system is able to adapt its automated data mining processes so that they are increasingly accurate over time.

This paper is structured as follows. Section 2 describes the conceptual framework of Cyclone and presents the proposed visual clustering based on spring forces. In Section 3 we introduce the adaptive visual clustering paradigm, which is capable of reacting to variations in users' categorization behavior and adjusting the forces on the

objects in real-time to reflect this. Section 4 presents initial results obtained from multi-user experiments. Finally, Section 5 concludes with a summary and future work.

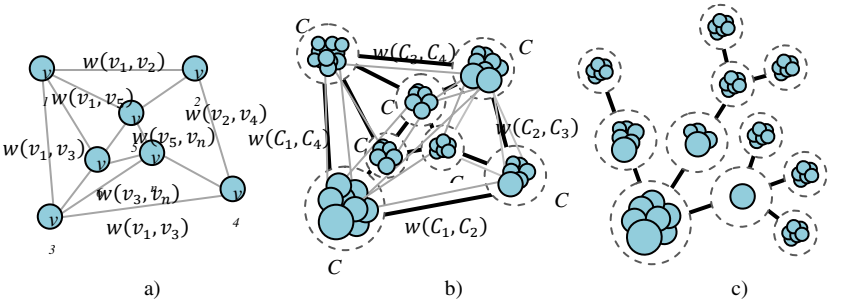

Fig. 1. The visual representations of a) an edge-weighted graph between visual objects, b) an edge-weighted graph between clusters and categories and c) a generated taxonomy T.

2 The Conceptual Framework

Cyclone is an intelligent agent-based visual framework offering a means for the user to exploit, analyze and categorize unstructured information from various sources into a more structured and manageable form [15]. The main strength of Cyclone is its capability to provide users with continuous visual feedback in real-time so as to increase their *understanding, confidence* and *trust* in the functional processes of the system, from clustering visualization to categorization.

The conceptual framework of Cyclone is based on the ideas provided by graph theory, which uses mathematical structures to model relations between a pair of visual objects, as well as visual data mining techniques, involving force-based real-time adaptive clustering and interface design. The adaptive learning system for clustering and automatic categorization is based on a single-layer feed-forward neural network utilizing a Hebbian rule style weight calculation.

The following subsections present and describe each of the aforementioned framework components.

2.1 Definitions

Definition 1. Let $G = (V, E)$ be a *graph of visual objects*, where V is the set of visual objects and E the set of edges, with $|V| = N$ and $|E| = M$. A visual object v_i has up to $N - 1$ edges (without self-loops) to visual objects of G. Each visual object v_i is equipped with a set of information, such as Name n, URL url, Description d as well as a set of metadata, i.e. tags t describing the content and context of the service or information it represents. An *edge-weighted graph* $G^* = (V, E, w)$ is a graph of objects, where V is the set of visual objects, E the set of edges and w the set of edge weights respectively. The edge weight is defined as $w: E \to \Re$, where $w \in [0,1]$ (see Fig. 1a).

Let $S \subseteq V$, the *subgraph* of G^* induced by S so that $G[S] = (S, E_s, w_S)$ where $E_s = \{e = (u, v) \in E \text{ such that } u \in S \text{ and } v \in E\}$. A subgraph S can also be regarded as a *cluster* or a *category* C_x.

Definition 2. Let $A = (V, w)$ be a *similarity matrix*, extracted from $G^* = (V, E, w)$, where V is the set of visual objects, w the set of weights of a pair of visual jects v_i, v_j, where if $i = j$ the $w(v_i, v_j) = 1$.

Likewise, let $A^* = (C, w)$ be a *cluster or category-based similarity matrix*, where C is the set of clusters/categories and w the set of weights of a pair of clusters/categories C_x, C_y where if $x = y$ the $w(C_x, C_y) = 1$ (see Fig. 1b).

Definition 3. A clustered graph (taxonomy) is an ordered quadruple $T = (V, C, A, A^*)$, where V defines the set of visual objects, C is the clusters set and A and A^* are the set of edges (and their corresponding weights) among the visual objects and clusters respectively (see Fig. 1c).

2.2 Visual Clustering

Cyclone's visual clustering component consists of the following steps [15]:

1. *Initializing the visual environment*.
   Cyclone loads the information and services from various given data sources and produces the initial display. The information is represented as a visual node and color-coded according to the data source it comes from. Cyclone offers a rich set of visualization and graph layout methods allowing the user to program and personalize their visual environment as seen fit.
2. *Calculating similarity among visual objects*
   Before Cyclone can start with the clustering process, it calculates the similarity among the visual objects based on the provided tag set. The cosine similarity function is used to obtain the degree of relevance between a pair of objects. The objects are considered identical if they both share the exact set of tags. The similarity and dissimilarity between objects in Cyclone is calculated using the following equation:

$$similarity\,(v_i, v_j) = \frac{\sum_{t_j \in T} v_i, v_j}{\sqrt{\sum_{t_j \in T} {v_i}^2} \sqrt{\sum_{t_j \in T} {v_j}^2}} \tag{1}$$

   Cyclone computes a similarity measurement for each pair of objects to obtain the similarity matrix *A* of the information space.
3. *Visual clustering based on Spring forces*
   Cyclone utilizes a force-directed layout to perform visual clustering where spring forces (simulated physical forces [16]) operate between the visual objects. Depending on the position and $similarity\,(v_i, v_j)$ of the visual objects, the nature of employed forces is determined, i.e. they may be either attractive or repulsive and to varying degrees. The strength of the forces for a pair of visual objects is derived from the similarity matrix *A*.

The following describes the forces applied to a pair of visual objects v_i and v_j (inter-cluster):

$$F_{v_{i,j}} = F^r_{v_{i,j}} + F^a_{v_{i,j}} \tag{2}$$

with

$$F^r_{v_{i,j}} = \sum_{j=1\ and\ j\neq i}^{N} (r - (similarity\ (v_i, v_j) * r) \tag{3}$$

and

$$F^a_{v_{i,j}} = \sum_{j=1\ and\ j\neq i}^{N} (similarity\ (v_i, v_j) * r) \tag{4}$$

where N is the total number of visual objects and r a predefined maximum radius of a pair of objects within G.

In addition to the inter-cluster forces, Cyclone defines the intra-cluster forces which act between groups of visual objects forming clusters, which is described as:

$$F^a_{C_{x,y}} = \sum_{x=1\ and\ y\neq x}^{N} similarity(\overline{C_x}, \overline{C_y}) * r \tag{5}$$

where N is the total number of clusters and $\overline{C_x}\ and\ \overline{C_y}$ the centre of the clusters.

The equilibrium state of the forces is found when

$$\sum_{i=1}^{N} F_i = \sum_{i=1}^{N} (F^r_i + F^a_i) = 0 \tag{6}$$

Once the forces are calculated, Cyclone assigns them to each visual object and executes them in a physical model, allowing the interactions of the forces to determine the resultant positions of the objects. The objects interact and eventually settle into a low-energy steady state representing an emergent clustering.

4. *Manual or Auto-Categorization*

Cyclone's categorization process involves the specification of a hierarchical system of categories (see Fig. 1c) as well as placing information into the nodes of this hierarchy to ultimately generate a personalized taxonomy. Cyclone offers two alternative modes: (1) *Manual Categorization*, where the user explicitly selects and assigns a subset of visual objects into categories and (2) *Auto-Categorization*, where Cyclone's intelligent agent learns the categorization habits of the user and automatically performs the categorization of their behalf. For the latter, the agent employs a single layer feed-forward neural network utilizing a hebbian style weight calculation, which creates a feedback loop between the user and the system. The agent maps tags to categories by continuously updating the edge weights each time the user categorizes a set of objects into a category. The weight increases if the co-occurrence of tags within a category increases and conversely, the weights decrease over time if previous categorization patterns are not repeated or are frequently changed. The two

modes of operation may be either used exclusively by users as they see fit or, more typically, in combination creating a mixed-initiative experience.

For a more detailed description on the intelligent automatic categorization learning process of Cyclone, please refer to [15].

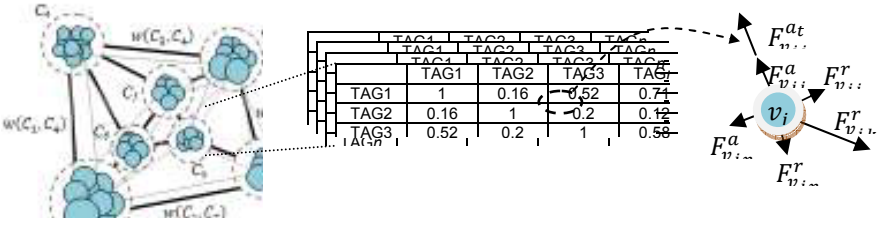

Fig. 2. The tag similarity calculation process for adaptive visual clustering

3 Adaptive Visual Clustering

A feedback loop between Cyclone and the user during the categorization process, taking a user's actions into account, updates the learnt model and visual clustering in an unobtrusive fashion. Adaptation of the model takes place each time the user assigns objects into categories and the visual clustering changes to reflect the categorization preferences of the user by updating and adjusting the forces on the visual objects.

The algorithm for the visual clustering adaptation is achieved by applying the following steps of *tag similarity calculation* (see Fig. 2) [15]:

- Step 1: Instantiate an edge-weighted graph for tags $G^* = (t, E, w)$ where t is the tags set, E the set of edges, and w the set of weights of the edges between a pair of tags. In addition, instantiate the similarity matrix $A^* = (t, w)$ based on G^*.
- Step 2: Set the edge weight $w_{i,j}$ for all possible pair of tags to zero, where $i \neq j$, and to 1, where $i = j$ (represents a self-loop which cannot be adjusted).
- Step 3: For categorization (manual or automatic), do the following:
  - Extract from the selected objects of the *same category* C_x the tags set $t_{C_x} \subseteq t$. For every tag combination t_a, t_b ($a \neq b$) of t_{C_i}, use the following equation to calculate the w_{t_a, t_b}:

$$(w_{t_a,t_b})^k = (w_{t_a,t_b})^{k-1}(1-\tau)\delta \tag{7}$$

  where $(w_{t_a,t_b})^{k-1}$ is the degree of similarity (weight) before applying Eq. 7 and τ is the decay value which is set to 0.05. The decay τ is an empirical value, which prevents the weights to increase endlessly and δ the pre-defined (and empirically selected) learning rate (0.1).
- Step 4: Adjust the strength of the forces for every visual object F_i^a using Eq. 5:

$$F_{v_{i,j}}^{a^*} = \sum_{j=1 \; and \; j \neq i}^{N} similarity\left(v_i, v_j, w\right) * r$$

where $similarity\left(v_i, v_j, w\right) = \frac{\sum_{t_j \in T} v_i, v_j, w_{t_a,t_b}}{\sqrt{\sum_{t_j \in T} v_i w_{t_a,t_b}}^2 \sqrt{\sum_{t_j \in T} v_j w_{t_a,t_b}}^2}$ (8)

The change in force strength causes the visual objects to rearrange and migrate to other positions so that the visual representations of the clusters adjust as a result.

4 Initial Experimental Results

We have conducted several multi-user experiments using real-data from different information sources and application context in order to validate the efficiency of Cyclone. Due to space constraints we focus on the qualitative aspects of the results in this paper.

The experiments were conducted on a group of researchers. Some of them were familiar with the task and others were non-experts and had no prior experience, however, neither group had used Cyclone before. Our hypothesis was that through using our proposed approach, in particular in its ability to visualize feedback in real-time, there wouldn't be a bias between the two groups in terms of their ability to familiarize themselves with the interface and use it efficiently to organize the data.

We chose a dataset consisting of 500 well-known songs (http://top500songs.blogspot.com) where each song was assigned 10 tags gleaned from the Audioscrobbler web service (http://www.audioscrobbler.net /data/webservices). A URL was also associated with each song such that the user could preview the song during the experiments.

The experiments were limited to 30 minutes, where the users were presented with the Cyclone interface (Fig. 3) showing a subset of the dataset read in and a visual clustering representation of it. The users were then asked to categorize the songs as they see fit through selecting and assigning of the visual objects.

An initial explanation stage was offered by the experimenter to ensure that the user was aware of the functional components of the interface. This stage, however, typically lasted only a couple of minutes as prescribed by how confident the users relayed they were with achieving their assigned task. It was observed that through successive categorizations by the users, the speed at which they categorized the objects increased. This effect is illustrated in Fig. 4 where the gradient of the curves gradually increase over time because of both an increased number of categorization actions, as well as a higher number of assignments per categorization.

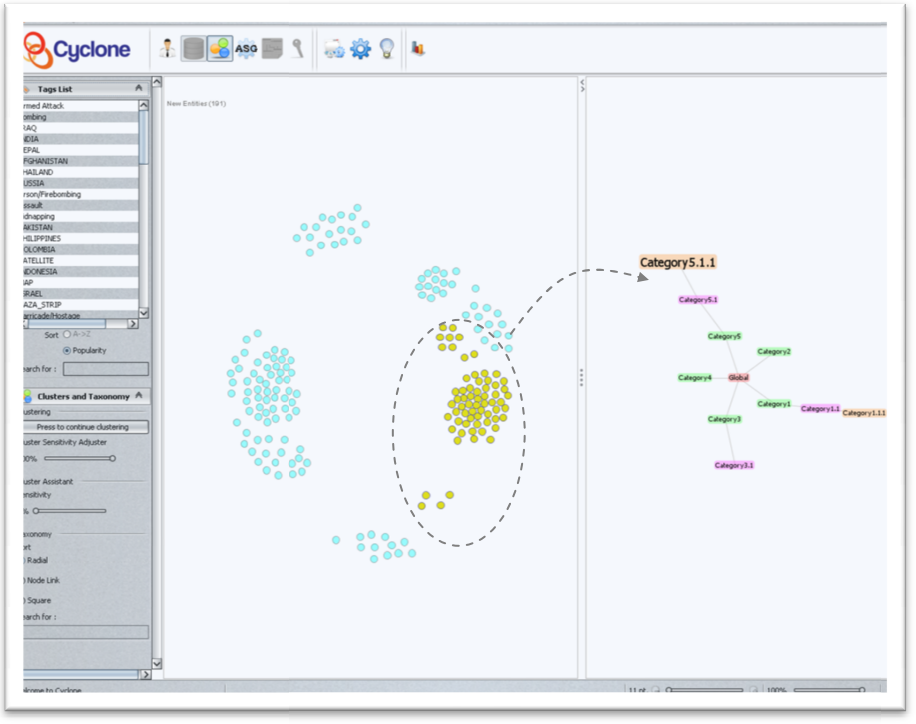

Fig. 3. The Cyclone Interface. It consists of three panels (incl. Tags Panel, Visual Clustering Panel and Taxonomy Panel) which the user can use to explore, analyze and exploit information. After the information are loaded (visual nodes), the user can initiate a categorization in a select-and-assign principle so that over time a personalized taxonomy is generated. Every time a categorization is performed and/or a change occurs, the visual clustering algorithm adjusts the force weights to adapt the visual clustering.

Clearly the effect above can be explained to an extent due to increased familiarization with the interface and dataset, but we argue that the speed at which this occurs is due in part by the real-time feedback offered by the interface. An important insight gained through observation and a post-experiment interviewing stage relates to the semantics of the categorization scheme chosen by users. A common attitude to dealing with categorizing the data was to, when encountering an object (song) which was deemed uninteresting, to assign it immediately to a category reflecting this. Unbeknown to the user at first, this behavior prompted the adaptive clustering to segregate those uninteresting objects from the interesting ones by visually migrating the objects on the screen. This allowed the users to concentrate more easily on the latter, creating more fine-grained categories for those objects which they were either more familiar with or valued more highly. This highlights the importance of the four primary goals for interactive data mining identified, but in particular the visualization of the system's interpretation of the data whilst incorporating users' preferences in real-time.

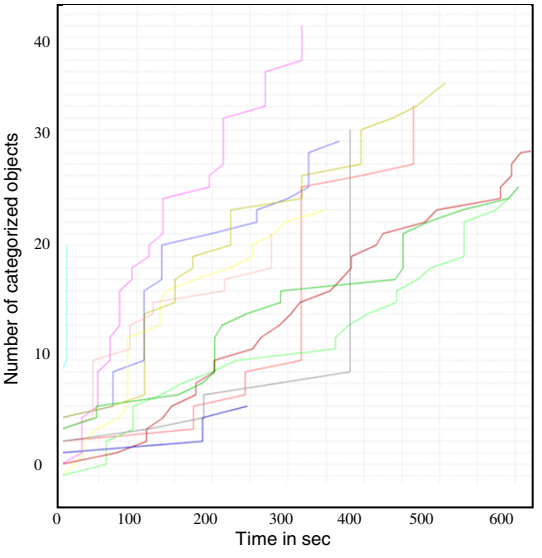

Fig. 4. User categorization process

5 Conclusions

The work presented discusses the case for introducing the provision to allow users to adapt the automated processes in an interactive data mining environment. In particular, we demonstrate the effectiveness of adaptive clustering to assist users and improve their ability to perform an information structuring task. Experimental investigation highlighted the importance of such real-time visual feedback to users and we believe that it is this visibility in particular which enables users of all skill levels to effectively tune the system to their benefit.

Current work involves applying the Cyclone system to various digital data sources such as Emails, web bookmarks, photo libraries and live video feeds to augment the information retrieval capabilities of users [17],[18].

Acknowledgments. The research described in this paper was conducted in the context of the Hyperion cluster project funded by the UK MoD Data and Information Fusion Defense Technology Centre (DIF DTC) research program and British Telecommunications plc (BT). We would also like to thank the members of the Centre for Information & Security Systems Research, in particular the Future Technologies Group at BT UK for their invaluable contributions and support, and also all the participants of the experiments presented in this paper for their involvement and constructive feedback.

References

1. Gants, J., Chute, C., Manfrediz, A., Minton, S., Reinsel, D., Schlichting, W., Toncheva, A.: The Diverse and Exploding Digital Universe – An Updated Forecase of Worldwide Information Growth Through 2011, IDC White Paper (2008)
2. Horvitz, E.: Principles of Mixed-Initiative User Interfaces. In: ACM conference on Human Factors in Computing Systems (CHI 1999), pp. 159–166 (1999)
3. Beale, R.: Supporting serendipity: Using ambient intelligence to augment user exploration for data mining and web browsing. J. of Human-Computer Studies 65, 421–433 (2007)
4. Ferreira de Oliveira, M.C., Levkowitz, H.: From Visual Data Exploration to Visual Data Mining: A Survey. IEEE Transactions on Visualization and Computer Graphics 9, 378–394 (2003)
5. Zhou, Y.: On Interactive Data Mining. In: Encyclopedia of Data Warehousing and Mining, vol. 2, pp. 1085–1090. Idea Group Inc. (2008)
6. Shneiderman, B.: Inventing Discovery Tools: Combining Information Visualization with Data Mining. Information Visualization 1, 5–12 (2002)
7. Xu, R., Wunsch II, D.: Survey of Clustering Algorithms. IEEE Transactions on Neural Networks 16, 645–678 (2005)
8. Gates, S.C., Teiken, W., Chen, K.-S.: Taxonomies by the numbers: building high-performance taxonomies. In: 14th ACM International Conference on Information and Knowledge Management, pp. 568–577 (2005)
9. Sanchez, D., Moreno, A.: Automatic Generation of Taxonomies from the WWW. In: Karagiannis, D., Reimer, U. (eds.) PAKM 2004. LNCS, vol. 3336, pp. 208–219. Springer, Heidelberg (2004)
10. Wetzker, R., Alpcan, T., Buckhage, C., Umbrath, W., Albayrak, S.: An unsupervised hierarchical approach to document categorization. In: 2007 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 482–486 (2007)
11. Lan, S.C., Al-Hawamdeh, S.: Taxonomy-Building Tools: An Investigative Study. J. Informatoin & Knowledge Management. 2, 63–77 (2003)
12. Wolin. B.: Method for Automatic Categorization of Items. US Patent 1 160 52 (2001)
13. Rostoker, C.: Interactive Visualization of the Market Graph. Technical report, University of British Columbia (2005)
14. DesJardins, M., Ferraioli, J.: Interactive Visual Clustering. In: 12th International Conference on Intelligent User Interfaces (2007)
15. Duman, H., Healing, A., Ghanea-Hercock, R.: An Intelligent Agent Approach for Visual Information Structure Generation. In: 2009 IEEE Symposium on Intelligent Agents, Nashville, Tennessee (2009)
16. Eades, P., Huang, M.L.: Navigating Clustered Graphs using Force-Directed Methods. J. Graph Algorithms and Applications 4, 157–181 (2000)
17. Ghanea-Hercock, R., Gelenbe, E., Jennings, N.R., Smith, O., Allsopp, D.N., Healing, A., Duman, H., Sparks, S., Karunatillake, N.C., Vytelingum, P.: Hyperion - Next Generation Battlespace Information Services. Computer Journal 50, 632–645 (2007)
18. Healing, A., Ghanea-Hercock, R., Duman, H., Jacob, M.: Nexus: Self-organising Agent-based Peer-to-Peer Middleware for Battlespace Support. Defence Industry Applications of Autonomous Agents and Multi-agent Systems, pp. 1–13 (2008)