ECAI 2012
Luc De Raedt et al. (Eds.)

© 2012 The Author(s).
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License.
doi:10.3233/978-1-61499-098-7-967

LSA for Mining Hidden Information in Action Game Semantics

Katia Lida Kermanidis and **Panagiotis Pandis** and **Costas Boletsis** and **Dimitra Chasanidou**[1]

Abstract. This paper describes the application of Latent Semantic Analysis to the term-document matrices that result from modeling an action game. Innovative solutions to address challenges like the definition of "words" and "documents" in the dynamic and complex domain of action games are proposed, and interesting, previously unknown semantic information is extracted.

1 INTRODUCTION

Latent Semantic Analysis (LSA) models documents and terms (words) in Information Retrieval in a way that allows the revelation of hidden underlying semantic relations between them that are not apparent at first sight [1], by dimensionality reduction (singular value decomposition) of the term-document matrix. In recent years its applicability has been extended to model the semantic domain of games, either board-like [2], or more complex dynamic environments [3]. Modeling the semantics of games allows for user-centered and intelligent game design.

Board-like games are more straightforward to model, as "words" (atomic semantic units that model a game state uniquely) and "documents" ("word" sequences that form a meaningful "utterance", i.e. a game session) are easily defined. In action games, the identification of the vocabulary and the utterances is more challenging as they are complex dynamic environments that are governed by causality, time-dependence and a set of relations among all entities, all of which are not obvious at first sight.

This paper describes the effect of applying LSA to the action game SpaceDebris [4] for player modeling, i.e. grouping players with similar gaming techniques together. Identifying the players' gaming techniques enables the design of games that are adaptable to the players' needs and individual style, and therefore more enjoyable [5].

2 MODELING SPACEDEBRIS

SpaceDebris concerns space battles with the player trying to destroy as many enemy spaceships as possible with his laser gun, and survive. Floating asteroids may indirectly be used to destroy enemy spaceships, shield and life power-ups are another indirect way for the player to strengthen his status. A screenshot of the game can be seen in Figure 1.

Two ways for representing "words" have been adopted. In the "holistic" representation, a "word" consists completely of non-spatial (distributed) information, e.g. score, number of available life upgrades, number of available shield upgrades, number of enemy ships close to the player, number of enemy ships very close to the player etc. (22 features in total). Discretization of the numeric features has been applied. In the "grid" representation the game terrain is viewed as an 11(rows)x8(columns) grid of cells. A "word" consists of two parts: the first denotes the concatenation of the states of all 88 cells, the second denotes out-of-the-grid (not spatially distributed) information, i.e. the score and the number of life and shield upgrades (91 features in total). There are 25 distinct cell states (empty cell, cell with asteroid, player ship, enemy ship, laser, shield upgrade, life upgrade, hit enemy ship, hit asteroid etc).

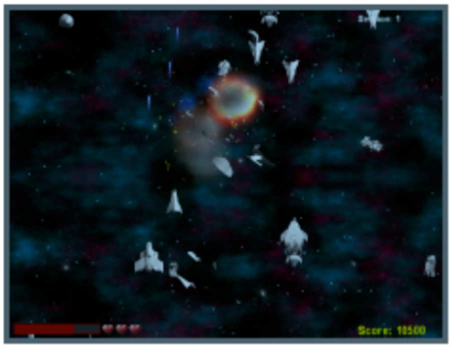

Figure 1. SpaceDebris

The game state ("word") is recorded every 0.5 seconds. Consecutive game states from the beginning until the end of a game session form the meaningful "utterance" of a player. The resulting words and game sessions are used to form the term-document matrix, the contents of which are the raw frequencies of each "word" in each session. The holistic and grid term-document matrices are very sparse, due to the large number of features, and the large number of distinct cell states.

3 EXPERIMENTAL SETUP

Player techniques are predefined: aggressive (a player keen on action games and when playing SpaceDebris fires constantly without frequent use of the power-ups), defensive a player keen on puzzle and internet games and when playing SpaceDebris does not fire or tries to avoid the enemies in order not to be killed, tactical (a player keen on playing strategy or adventure games and when playing SpaceDebris makes wise use of the laser and power-ups) and novice (a player with little gaming experience and playing SpaceDebris without any particular style).

[1] Department of Informatics, Ionian University, Corfu, Greece, email: kerman@ionio.gr

The participants included 10 players (74 game sessions, 10532 game states). Each game state constitutes a learning vector in the dataset. Each player is assigned a technique based on the observations of experts on the player's game during a trial gaming period. 29% of the vectors belong to the novice class, 42% to the tactical, 19% to the aggressive and 10% to the defensive class.

Classification (C4.5) was applied first for classifying game states to one of the four styles. Experiments were run using 10-fold cross validation. Classification precision and recall are shown in Table 1. Lower results for the defensive class may be attributed to a large degree to its rare occurrence in the data, compared to the other classes.

Table 1. Classification results.

	C4.5 (holistic)		C4.5 (grid)	
	Pr	Re	Pr	Re
Aggressive	0.64	0.61	0.55	0.52
Tactical	0.72	0.74	0.54	0.55
Defensive	0.57	0.58	0.47	0.31
Novice	0.62	0.6	0.5	0.56

The non-distributed (holistic) results in the present approach are higher than the distributed ones (grid), due to the pre-processing (discretization) of the numeric features of the holistic dataset.

The relatively arbitrary (not fully objective or unambiguous) manner of assigning style tags to players is one of the two main reasons for experimenting with unsupervised learning. The other reason is that clustering similar gaming styles together may reveal hidden, previously unknown information regarding the data. *K*-means is used for clustering and the number of clusters is initially set to 4, in order to enable clusters-to-classes evaluation. Results are decent but not that exciting, as can be seen in Figure 2.

```
   0    1    2    3  <-- assigned to cluster
1671 1460  924 1385 | Tactical
 105  860  971  104 | Aggressive
  68  169  112 1652 | Novice
   0    0    0 1051 | Defensive
```

Figure 2. Clusters-to classes evaluation on the holistic dataset.

LSA is performed on the holistic and the grid term-document matrices. Several experiments were run with various dimensionality reduction ratios (i.e. maintaining 20, 40 and 60 singular values). Clustering is performed on the resulting V matrices (the matrices that encode the transformation of the game sessions into the 20, 40 or 60 latent semantic space dimensions).

3.1 Revealing hidden information

An interesting observation becomes apparent when the number of clusters is set to 2. Using the initial (prior to LSA) datasets k-means groups the instances into two classes with no apparent relation to the four styles. After performing LSA, however, the vast majority of the instances of the novice and defensive classes are grouped together and form one cluster, while instances of the tactical and aggressive classes form the second cluster. Tables 2 and 3 show the number of outliers for the two clusters for the holistic and the grid datasets respectively divided by the total number of members assigned to the cluster. Cluster 1 outliers are the defensive and novice instances that are grouped into cluster 2. Cluster 2 outliers are the aggressive and tactical instances that are grouped into cluster 1. The two formed clusters are interesting and can be explained, as novice players usually tend to play defensively, with no pattern or offensive strategy. On the other hand, aggressive and tactical players share the same confidence and a strategic plan to win. LSA revealed these previously unknown relations.

Table 2. Number of outliers – holistic dataset.

	Singular values		
	20	40	60
Cluster 1 (Defensive and Novice)	3/40	1/34	13/25
Cluster 2 (Aggressive and Tactical)	0/34	4/40	1/49

Table 3. Number of outliers – grid dataset.

	Singular values		
	20	40	60
Cluster 1 (Defensive and Novice)	4/40	4/37	4/40
Cluster 2 (Aggressive and Tactical)	1/34	4/37	1/34

The distributed modeling (the grid) seems to affect positively clustering performance as the number of singular values increases.

CONCLUSION

In this paper we described the effect of LSA on modeling the semantic space of action videogames with the ultimate goal to model the players' gaming style. Two modeling schemata were adopted, one non-spatially-distributed ("holistic") and one spatially-distributed ("grid"). LSA manages to reveal previously unknown, hidden semantic relations among the data instances.

Several future research directions are worth exploring. Instead of using raw term frequencies in the term-document matrices, other weights (e.g. td-idf) would be interesting to experiment with. Classification and clustering based on the individual player, and not the player's style, could also constitute a challenging perspective that might reveal further interesting semantic information hidden in the data. Finally, the sparseness of the term-document matrices could be addressed by performing feature selection (thereby reducing the number of features) or by reducing the number of distinct cell states in the grid (e.g. by merging together states that may be considered equivalent in modeling the game space).

REFERENCES

[1] T. Landauer, P. Foltz, and D. Laham, 'An introduction to latent semantic analysis', *Discourse Processes*, 25, 259-284, (1998).

[2] B. Lemaire, 'Models of high-dimensional semantic spaces', *4th International Workshop on Multistrategy Learning,* (1998).

[3] J.F. Quesada, W. Kintsch, and E. Gomez, 'A computational theory of complex problem solving using the vector space model (part I): latent semantic analysis, through the path of thousands of ants', In J. J. Cañas (Ed.), *Cognitive Research with Microworlds*, 117-131, Granada, Spain, (2001).

[4] K. Anagnostou, and M. Maragoudakis, 'Data mining for player modeling in videogames', *Panhellenic Conference on Informatics*, 30-34, (2009).

[5] C. Boletsis, D. Chasanidou, P.Pandis and K.L.Kermanidis, 'Semantic representation of action games', *Workshop on Machine Learning and Data Mining in Games,* (2011).