Sentic Medoids: Organizing Affective Common Sense Knowledge in a Multi-Dimensional Vector Space

Erik Cambria[1], Thomas Mazzocco[1], Amir Hussain[1], and Chris Eckl[2]

[1] Cognitive Signal and Image Processing Research Laboratory
University of Stirling, Stirling, UK
[2] Sitekit Labs, Sitekit Solutions Ltd, Portree, UK
{eca,tma,ahu}@cs.stir.ac.uk, chris.eckl@sitekit.net
http://cs.stir.ac.uk/~eca/sentics

Abstract. Existing approaches to opinion mining and sentiment analysis mainly rely on parts of text in which opinions and sentiments are explicitly expressed such as polarity terms and affect words. However, opinions and sentiments are often conveyed implicitly through context and domain dependent concepts, which make purely syntactical approaches ineffective. To overcome this problem, we have recently proposed Sentic Computing, a multi-disciplinary approach to opinion mining and sentiment analysis that exploits both computer and social sciences to better recognize and process opinions and sentiments over the Web. Among other tools, Sentic Computing includes AffectiveSpace, a language visualization system that transforms natural language from a linguistic form into a multi-dimensional space. In this work, we present a new technique to better cluster this vector space and, hence, better organize and reason on the affective common sense knowledge in it contained.

Keywords: Sentic Computing, AI, Semantic Web, NLP, Clustering, Opinion Mining and Sentiment Analysis.

1 Introduction

The ways people express their opinions and sentiments have radically changed in the past few years thanks to the advent of social networks, online communities, blogs, wikis and other online collaborative media. But, although these online social data are perfectly suitable for human consumption, they remain hardly accessible to machines.

To bridge the cognitive and affective gap between word-level natural language data and the concept-level opinions and sentiments conveyed by them, we recently developed Sentic Computing [1], a multi-disciplinary approach to opinion mining and sentiment analysis that exploits both computer and social sciences to better recognize, interpret and process opinions and sentiments over the Web.

The main tool used, within Sentic Computing, to perform emotive reasoning on natural language is AffectiveSpace [2], a multi-dimensional representation of

D. Liu et al. (Eds.): ISNN 2011, Part III, LNCS 6677, pp. 601–610, 2011.

affective common sense knowledge. So far, a k-nearest neighbor (k-NN) approach has been used to divide this vector space into emotional clusters but resulting classification has tended to be imprecise for some less common affective concepts, which has often reflected on the overall analysis of texts, both for sentiment inference and polarity detection tasks.

To this end we developed a new technique, which we call sentic medoids, which more conveniently organizes AffectiveSpace by iteratively selecting appropriate concepts in order to minimize the sum of dissimilarities between them and the other concepts within the same cluster.

The structure of the paper is the following: Section 2 presents the main existing approaches to opinion mining and sentiment analysis, Section 3 explains the AffectiveSpace process, Section 4 illustrates the emotion categorization model adopted, Section 5 explains in detail the new technique developed for better clustering AffectiveSpace, Section 6 presents an evaluation of the technique and Section 7 comprises concluding remarks and a description of future work.

2 Opinion Mining and Sentiment Analysis

Existing approaches to automatic identification and extraction of opinions and sentiments from text can be grouped into four main categories: keyword spotting, lexical affinity, statistical methods and Sentic Computing.

In keyword spotting [3][4][5] text is classified into categories based on the presence of fairly unambiguous affect words. Lexical affinity [6][7] assigns arbitrary words a probabilistic affinity for a particular opinion or emotion. Statistical methods [8][9] calculate the valence of keywords, punctuation and word co-occurrence frequencies on the base of a large training corpus. In Sentic Computing, whose term derives from the Latin 'sentire' (the root of words such as sentiment and sensation) and 'sensus' (intended as common sense), the analysis of text is not based on statistical learning models but rather on common sense reasoning tools [10] and domain-specific ontologies [11].

While the first three approaches mainly rely on parts of text in which opinions and sentiments are explicitly expressed such as polarity terms (e.g. good, bad, nice, nasty, excellent, poor) and affect words (e.g. happy, sad, calm, angry, interested, bored), Sentic Computing analyzes text at semantic-level, aiming to infer also the opinions and sentiments implicitly conveyed.

Moreover, unlike statistical classification (which generally requires large inputs and thus cannot appraise texts with satisfactory granularity), Sentic Computing enables the analysis of documents not only on the page- or paragraph-level but also on the sentence-level.

3 AffectiveSpace

AffectiveSpace is built by blending ConceptNet [12], a directed graph representation of common sense knowledge, with WordNet-Affect (WNA) [13], a linguistic

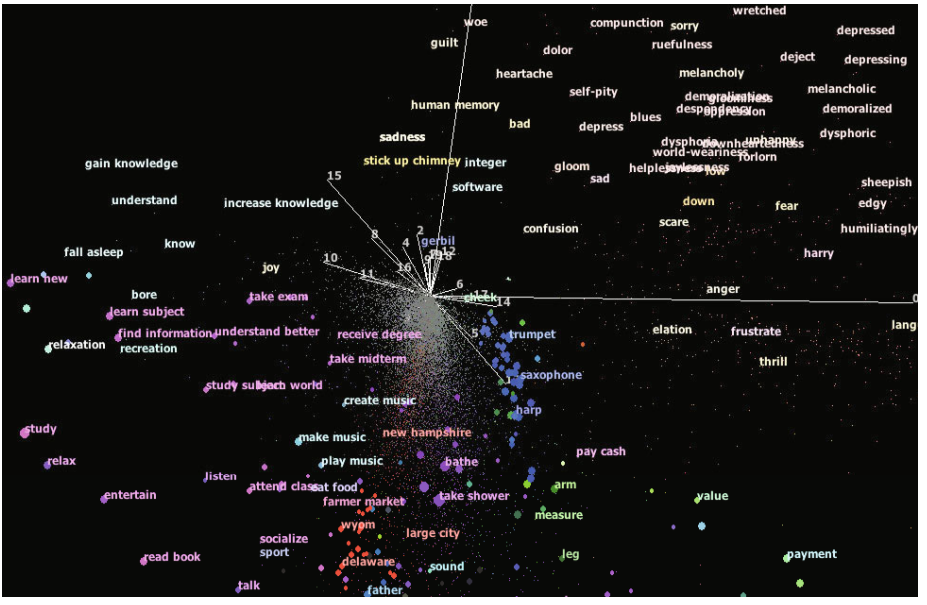


Fig. 1. AffectiveSpace

resource for the lexical representation of affective knowledge (Fig. 1). This alignment operation yields *AffectNet*: a new dataset in which common sense and affective knowledge coexist i.e. a matrix 14,301 × 117,365 whose rows are concepts (e.g. 'dog' or 'bake cake'), whose columns are either common sense and affective features (e.g. 'isA-pet' or 'hasEmotion-joy'), and whose values indicate truth values of assertions. Therefore, in *AffectNet*, each concept is represented by a vector in the space of possible features whose values are positive for features that produce an assertion of positive valence (e.g. 'a penguin is a bird'), negative for features that produce an assertion of negative valence (e.g. 'a penguin cannot fly') and zero when nothing is known about the assertion.

The degree of similarity between two concepts, then, is the dot product between their rows in *AffectNet*. The value of such a dot product increases whenever two concepts are described with the same feature and decreases when they are described by features that are negations of each other. When performed on *AffectNet*, however, these dot products have very high dimensionality and are difficult to work with. In order to approximate them in a useful way, we project all of the concepts from the space of features into a space with many fewer dimensions i.e. we reduce the dimensionality of *AffectNet*.

In particular, we perform truncated singular value decomposition (TSVD) [14] on *AffectNet* and obtain a new matrix, *AffectNet**, which forms a low-rank approximation of the original data. This estimation is based on minimizing the Frobenius norm of the difference between *AffectNet* and *AffectNet** under the constraint $rank(AffectNet^*) = k$ and it represents the best approximation of

AffectNet in the least-square sense (for the Eckart–Young theorem [15]). In particular, we choose to discard all but the first 100 principal components and hence obtain AffectiveSpace, a 100-dimensional space in which different vectors represent different ways of making binary distinctions among concepts and emotions.

In AffectiveSpace common sense and affective knowledge are in fact combined, not just concomitant, i.e. everyday life concepts like 'have breakfast', 'meet people' or 'watch tv' are linked to a hierarchy of affective domain labels. By exploiting the information sharing property of TSVD, concepts with the same affective valence are likely to have similar features i.e. concepts concerning the same opinion tend to fall near each other in the vector space.

Concepts and emotions are represented by vectors of 100 coordinates: these coordinates can be seen as describing concepts in terms of 'eigenmoods' that form the axes of AffectiveSpace i.e. the basis e_0,...,e_{99} of the vector space. For example, the most significant eigenmood, e_0, represents concepts with positive affective valence. That is, the larger a concept's component in the e_0 direction is, the more affectively positive it is likely to be. Consequently concepts with negative e_0 components have negative affective valence.

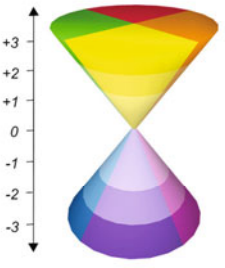
4 The Hourglass of Emotions

To reason on the disposition of concepts in AffectiveSpace, we use the Hourglass of Emotions [16], a novel affective categorization model in which sentiments are organized around four independent – but concomitant – dimensions, whose different levels of activation make up the total emotional state of the mind.

The Hourglass of Emotions, in fact, is based on the idea that the mind is made of different independent resources and that emotional states result from turning some set of these resources on and turning another set of them off [17]. Each such selection changes how we think by changing our brain's activities: the state of 'anger', for example, appears to select a set of resources that help us react with more speed and strength while also suppressing some other resources that usually make us act prudently.

The primary quantity we can measure about an emotion we feel is its strength. But when we feel a strong emotion it is because we feel a very specific emotion. And, conversely, we cannot feel a specific emotion like 'fear' or 'amazement' without that emotion being reasonably strong. Mapping this space of possible emotions leads to an hourglass shape (Fig. 2). In the Hourglass of Emotions, affective states are not classified, as often happens in the field of emotion analysis, into basic emotional categories, but rather into four independent and concomitant dimensions – Pleasantness, Attention, Sensitivity and Aptitude – in order to understand how much respectively:

1. the user is amused by interaction modalities (Pleasantness)
2. the user is interested in interaction contents (Attention)
3. the user is comfortable with interaction dynamics (Sensitivity)
4. the user is confident in interaction benefits (Aptitude)



	Pleasantness	Attention	Sensitivity	Aptitude
+3	ecstasy	vigilance	rage	admiration
+2	joy	anticipation	anger	trust
+1	serenity	interest	annoyance	acceptance
0	—	—	—	—
-1	pensiveness	distraction	apprehension	boredom
-2	sadness	surprise	fear	disgust
-3	grief	amazement	terror	loathing

Fig. 2. The Hourglass of Emotions

Each affective dimension is characterized by six levels of activation, called 'sentic levels', which determine the intensity of the expressed/perceived emotion as an $int \in [-3,3]$. These levels are also labeled as a set of 24 basic emotions (six for each of the affective dimensions) in a way that allows the model to specify the affective information associated with text both in a dimensional and in a discrete form. The dimensional form, in particular, is called 'sentic vector' and it is a four-dimensional $float$ vector that can potentially express any human emotion in terms of Pleasantness, Attention, Sensitivity and Aptitude.

Some particular sets of sentic vectors have special names as they specify well-known compound emotions. For example the set of sentic vectors with a level of Pleasantness $\in (1,2]$ ('joy'), a null Attention, a null Sensitivity and a level of Aptitude $\in (1,2]$ ('trust') are called 'love sentic vectors' since they specify the compound emotion of 'love'.

5 Sentic Medoids

So far, a k-NN approach has been used to divide AffectiveSpace into clusters according to the Hourglass model. In particular, the centroid of each cluster has been originally taken as the concept corresponding to each sentic level of the

described model. This choice, because of its simplicity, can pose a few limitations on the effectiveness of division into clusters of AffectiveSpace.

For example, at present, there are no evidences that the concepts corresponding to a specific sentic level is necessarily the best choice for clustering; moreover, it is not proved that a centroid must coincide with an actual concept instead of being any point in the space; further, this division of space leads inevitably to poor performances in classification of concepts with low affective valence. Within this study we address the first of these points, trying to provide with centroids, which may be regarded as better (in some sense) than the currently used ones.

K-means [18] and k-medoids [19] are two widely used techniques for clustering whose aim is to partition the given observations into k clusters around as many centroids, trying to minimize a given cost function. While k-means algorithm does not pose constraints on centroids, k-medoids do assume that centroids must coincide with k observed points. At this stage, a k-medoids approach has been chosen to investigate if the concepts that are currently used as centroids need to be replaced with different ones.

The most commonly used algorithm for finding the k medoids is the partitioning around medoids (PAM) algorithm. The PAM algorithm determines a medoid for each cluster selecting the most centrally located centroid within the cluster. After selection of medoids, clusters are rearranged so that each point is grouped with the closest medoid. Since k-medoids clustering is a NP-hard problem [20], different approaches based on alternative optimization algorithms have been developed, though taking risk of being trapped around local minima.

For the purpose of this work, we use a modified version of the algorithm recently proposed by Park and Jun [21], which runs similarly to the k-means clustering algorithm. This has shown to have similar performance when compared to PAM algorithm while taking a significantly reduced computational time. In particular, we have N concepts ($N = 14,301$) encoded as points $x \in \mathbb{R}^p (p = 50)$. We want to group them into k clusters and, in our case, we can fix $k = 24$ as we are looking for one cluster for each sentic level s of the Hourglass model.

Generally, the initialization of clusters for clustering algorithms is a problematic task as the process often risks to get stuck into local optimum points, depending on the initial choice of centroids [22]. However, for this study, we decide to use as initial centroids the concepts which are currently used as centroids for clusters, as they specify the emotional categories we want to organize AffectiveSpace into. For this reason, what is usually seen as a limitation of the algorithm can be seen as advantage for this study, since we are not looking for the 24 centroids leading to the best 24 clusters but indeed for the 24 centroids identifying the required 24 sentic levels (i.e. the centroids should not be 'too far' from the ones currently used).

In particular, as the Hourglass affective dimensions are independent but concomitant, we need to cluster AffectiveSpace four times, once for each dimension. According to the Hourglass categorization model, in fact, each concept can convey, at the same time, more than one emotion (which is why we get compound emotions) and this information can be expressed via a sentic vector specifying the

concept's affective valence in terms of Pleasantness, Attention, Sensitivity and Aptitude. Therefore, given that the distance between two points in AffectiveSpace is defined as $D(a,b) = \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2}$ (note that the choice of Euclidean distance is arbitrary), the used algorithm, applied for each of the four affective dimensions, can be summarized as follows:

1. Each centroid $C_n \in \mathbb{R}^{50}\,(n = 1, 2, ..., k)$ is set as one of the six concepts corresponding to each s in the current affective dimension.
2. Assign each record x to a cluster Ξ so that $x_i \in \Xi_n$ if $D(x_i, C_n) \leq D(x_i, C_m)$ $m = 1, 2, ..., k$.
3. Find a new centroid C for each cluster Ξ so that $C_j = x_i$ if $\sum_{x_m \in \Xi_j} D(x_i, x_m)$ $\leq \sum_{x_m \in \Xi_j} D(x_h, x_m) \quad \forall x_h \in \Xi_j$.
4. Repeat step 2 and 3 until no changes on centroids are observed.

Note that condition posed on steps 2 and 3 may occasionally lead to more than one solution. Should this happen, our model will randomly choose one of them. This clusterization of AffectiveSpace allows to calculate, for each common sense concept x, a four-dimensional sentic vector that defines its affective valence in terms of a degree of fitness $\mathbf{f}(x)$ where $f_a = D(x, C_j) \;\; C_j | D(x, C_j) \leq D(x, C_k)$ $a = 1, 2, 3, 4 \quad k = 6a\text{-}5, 6a\text{-}4, ..., 6a$.

6 Evaluation

In order to evaluate the proposed technique, we exploited a corpus of 5,000 mood-tagged blogs from LiveJournal (LJ) [23], a virtual community that allow users to keep a blog, journal or diary and to label their posts with a mood label, by choosing from more than 130 predefined moods or by creating custom mood themes. In particular, we both embedded sentic medoids in the sentics extraction process (Fig. 3), to test the technique on natural language data, and built a benchmark for affective common sense knowledge (BACK), in order to more specifically calculate statistical classifications such as precision and recall.

We compared the outputs of sentics extraction process, using a k-NN approach for clustering AffectiveSpace firstly and employing sentic medoids secondly. Results showed the latter technique to be up to 16% more accurate than the former. Classification of 'happy' and 'sad' posts, for example, was performed with 89% and 81% precision respectively and recall rates of 77% and 73% using sentic medoids, while k-NN identified the same posts with 75% and 62% precision and 70% and 58% recall respectively. The F-measure values obtained, hence, were significantly better: 82% and 76% for 'happy' and 'sad' posts using sentic medoids versus 72% and 60% respectively using k-NN, for an average accuracy improvement of 13%.

To build BACK we applied CF-IOF (concept frequency - inverse opinion frequency) [24] on the LJ corpus. CF-IOF is a technique that identifies common domain-dependent semantics in order to evaluate how important a concept is to

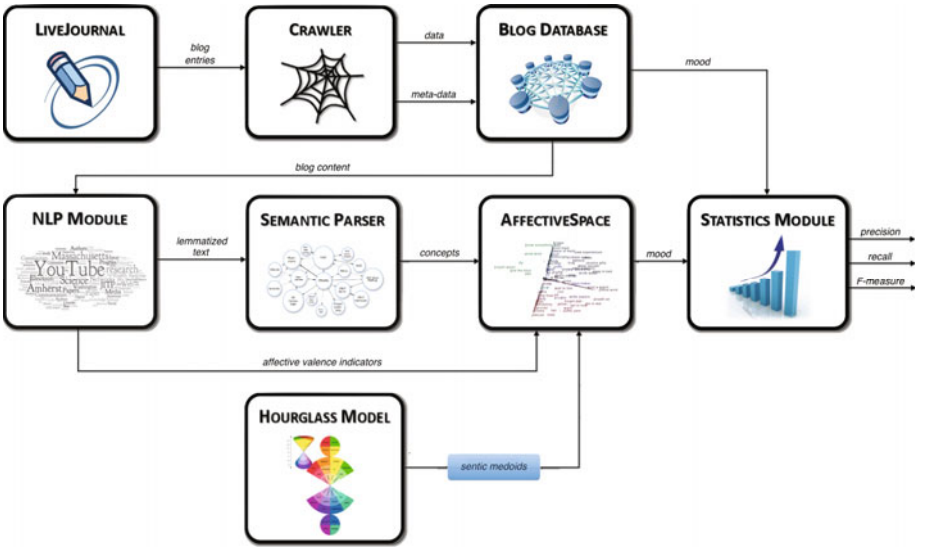


Fig. 3. Evaluation of sentics extraction process

a set of opinions concerning the same topic. Firstly, the frequency of a concept c for a given domain d is calculated by counting the occurrences of the concept c in the set of available d-tagged opinions and dividing the result by the sum of number of occurrences of all concepts in the set of opinions concerning d. This frequency is then multiplied by the logarithm of the inverse frequency of the concept in the whole collection of opinions, that is:

$$CF\text{-}IOF_{c,d} = \frac{n_{c,d}}{\sum_k n_{k,d}} \log \sum_k \frac{n_k}{n_c}$$

where $n_{c,d}$ is the number of occurrences of concept c in the set of opinions tagged as d, n_k is the total number of concept occurrences and n_c is the number of occurrences of c in the whole set of opinions. A high weight in CF-IOF is reached by a high concept frequency in a given domain and a low frequency of the concept in the whole collection of opinions.

We exploited CF-IOF weighting to filter out common concepts in the LJ corpus and detect relevant mood-dependent semantics for each of the Hourglass sentic levels. The result was a benchmark of 2000 affective concepts that we compared with the classification results obtained by applying k-NN and sentic medoids on AffectiveSpace. Also in this case, the latter technique outperformed the former. In particular, the average precision gap between sentic medoids and k-NN was +9% and the mean recall difference +15%.

7 Conclusion and Future Efforts

In a world in which millions of people express their opinions and sentiments in blogs, wikis, fora, chats and social networks, the distillation of knowledge from this huge amount of unstructured information is a very arduous task.

While existing approaches mainly work at syntactic-level, we use Sentic Computing techniques and tools to analyze natural language text at semantic-level. In particular, we infer opinions and sentiments from the Web using AffectiveSpace, a multi-dimensional vector space of affective common sense knowledge.

In this work, we proposed a new technique to cluster this vector space according to a novel emotion categorization model based on the idea that the mind is made of different independent resources and that emotional states result from turning some set of these resources on and turning another set of them off.

In particular, sentic medoids organize the affective common sense knowledge contained in AffectiveSpace by iteratively selecting appropriate concepts in order to minimize the sum of dissimilarities between them and the other concepts within the same cluster, and, hence, lead to more accurate results for sentiment inference and polarity detection tasks.

Whilst this study has shown encouraging results, further research studies are now planned to investigate if and how the centroids selection could be still improved. In particular, we would state whether modifications of the dissimilarity function and/or of the centroids' constraints could lead to an improvement of the model performance with special regard to concepts with low affective valence.

Soon, we will extend BACK using a bigger LJ corpus and make it publicly available for comparing Sentic Computing with other opinion mining and sentiment analysis techniques. Moreover, we plan to exploit BACK to make a comprehensive comparison between the proposed method and other popular clustering techniques such as maximum entropy classifiers, naïve Bayes classifiers, support vector machines (SVMs) and neural networks, in order to select the best method for organizing affective common sense knowledge and, ultimately, improve Sentic Computing.

Acknowledgements

This research has been funded by the UK Engineering and Physical Sciences Research Council (EPSRC Grant Reference No. EP/G501750/1) and Sitekit Labs. This work was undertaken during the first author's research visit to the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese of Academy of Sciences (CAS) in Beijing (China), which was jointly funded by the Royal Society of Edinburgh (UK) and CAS.

References

1. Cambria, E., Hussain, A., Havasi, C., Eckl, C.: Sentic Computing: Exploitation of Common Sense for the Development of Emotion-Sensitive Systems. In: Esposito, A., Campbell, N., Vogel, C., Hussain, A., Nijholt, A. (eds.) Second COST 2102. LNCS, vol. 5967, pp. 153–161. Springer, Heidelberg (2010)

2. Cambria, E., Hussain, A., Havasi, C., Eckl, C.: AffectiveSpace: Blending Common Sense and Affective Knowledge to Perform Emotive Reasoning. In: WOMSA at CAEPIA, Seville (2009)
3. Elliott, C.: The Affective Reasoner: A Process Model of Emotions in a Multi-agent System. The Institute for the Learning Sciences, Technical Report No. 32 (1992)
4. Ortony, A., Clore, G., Collins, A.: The Cognitive Structure of Emotions. Cambridge University Press, New York (1988)
5. Wiebe, J., Wilson, T., Claire, C.: Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation 39(3), 165–210 (2005)
6. Somasundaran, S., Wiebe, J., Ruppenhofer, J.: Discourse Level Opinion Interpretation. In: COLING, Manchester (2008)
7. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In: HLT-EMNLP, Vancouver (2005)
8. Hu, M., Liu, B.: Mining Opinion Features in Customer Reviews. In: AAAI, San Jose (2004)
9. Goertzel, B., Silverman, K., Hartley, C., Bugaj, S., Ross, M.: The Baby Webmind Project. In: AISB, Birmingham (2000)
10. Cambria, E., Hussain, A., Havasi, C., Eckl, C.: Common Sense Computing: From the Society of Mind to Digital Intuition and beyond. In: Fierrez, J., Ortega-Garcia, J., Esposito, A., Drygajlo, A., Faundez-Zanuy, M. (eds.) BioID MultiComm2009. LNCS, vol. 5707, pp. 252–259. Springer, Heidelberg (2009)
11. Cambria, E., Grassi, M., Hussain, A., Havasi, C.: Sentic Computing for Social Media Marketing. To appear in: Multimedia Tools and Applications. Springer, Heidelberg (2010)
12. Havasi, C., Speer, R., Alonso, J.: ConceptNet 3: a Flexible, Multilingual Semantic Network for Common Sense Knowledge. In: RANLP, Borovets (2007)
13. Strapparava, C., Valitutti, A.: WordNet-Affect: an Affective Extension of WordNet. In: LREC, Lisbon (2004)
14. Wall, M., Rechtsteiner, A., Rocha, L.: Singular Value Decomposition and Principal Component Analysis. In: Berrar, D., et al. (eds.) A Practical Approach to Microarray Data Analysis, pp. 91–109. Kluwer, Norwell (2003)
15. Eckart, C., Young, G.: The Approximation of One Matrix by Another of Lower Rank. Psychometrika 1(3), 211–218 (1936)
16. Cambria, E., Hussain, A., Havasi, C.: Eckl: SenticSpace: Visualizing Opinions and Sentiments in a Multi-Dimensional Vector Space. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010. LNCS (LNAI), vol. 6279, pp. 385–393. Springer, Heidelberg (2010)
17. Minsky, M.: The Emotion Machine. Simon and Schuster, New York (2006)
18. Hartigan, J., Wong, M.: Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society 28(1): 100-108 (1979)
19. Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
20. Garey, M., Johnson, D.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman, New York (1979)
21. Park, H., Jun, C.: A Simple and Fast Algorithm for K-Medoids Clustering. Expert System with Applications 36, 3336–3341 (2009)
22. Duda, R., Hart, P.: Pattern Classification and Scene Analysis. Wiley, New York (1973)
23. LiveJournal, `http://www.livejournal.com`
24. Cambria, E., Hussain, A., Durrani, T., Havasi, C., Eckl, C., Munro, J.: Sentic Computing for Patient Centered Applications. In: IEEE ICSP 2010, Beijing (2010)