Onomatology and content analysis of ergodic literature

Eugenia-Maria Kontopoulou
CEID, University of Patras
Rio, Greece
kontopoulo@ceid.upatras.gr

Maria Predari
CEID, University of Patras
Rio, Greece
predari@ceid.upatras.gr

Efstratios Gallopoulos
CEID, University of Patras
Rio, Greece
stratis@ceid.upatras.gr

ABSTRACT

We first establish a connection between the concept of ergodicity in mathematics and "ergodic literature" of the Choose-your-own-Adventure (CYOA) type that serves to answer some existing objections regarding the use of the term in the latter context. We then consider some steps towards the construction of concept maps for CYOA-type ergodic literature. Our analysis is based on modeling ergodic literature using digraphs and matrices. Promising preliminary results are obtained using content to refine link-based ranking.

Categories and Subject Descriptors

H.5.4 [**Information Systems**]: Hypertext/Hypermedia—*Theory*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing Methods,Linguistic Processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Clustering*; J.5 [**Computer Applications**]: Arts and Humanities—*Language translation,Linguistics,Literature*

General Terms

Algorithms, Experimentation, Measurement, Clustering

Keywords

hypertext, interactive fiction, Choose Your Own Adventure fiction, storylet, ergodic literature, Markov chains, ergodic theory, link-based ranking, digraph, stochastic matrix, PageRank, TMG, Evgenios Trivizas

1. INTRODUCTION

If "literature in the twenty-first century is computational" ([14]) it is natural to seek the application of mathematical techniques, algorithms and the computer to analyze it[1]. In [16] we attempted this in order to uncover latent properties of ergodic texts. The investigation was based on two fundamental mathematical structures: The Graph and the Matrix, applied to children's books of the CYOA ("Choose Your Own Adventure") type. Moreover, we considered algorithms for ranking storylets (the "special purpose" lexias in CYOA literature). These relied strictly on the link structure of these books to induce a ranking and were based on methods such as PageRank. We conjectured and provided some supporting evidence that the ranking was useful for constructing a "concept map" (CM) that we view as "information providers" for the multiple plots that coexist within each book.

In this "work-in-progress" paper we first establish a connection that can be used in support of the use of the term "ergodic" for CYOA type literature and also make some remarks on terminology. We then present preliminary results from the clustering of the possible plots within the existing link structure and show that the information mined can help to refine the CMs obtained from the link-based ranking. Experiments were conducted using data from the CHILDIF collection that was created for our study in [16]. Twelve books were used from the following series: 1) "Choose Your Own Adventure", published by CYOA corp., 2) "Innerstar University", published by American Girl and 3) "Multiclone Tales", published by Kalendis (in Greek).

2. ERGODIC ONOMATOLOGY

We recall from [2] that "ergodic literature" refers to the genre of books in which "there is non-trival effort to traverse the text" (they are demanding, according to [4]).

In the course of our investigations, we found that the vast majority of storylets from the CYOA and American Girl series had only 1, 2 and very rarely 3 outgoing links (that is that the nodes of the corresponding digraph had out-degree 1 or 2 and occasionally 3). The books from the third series (authored by Evgenios Trivizas) are even more demanding, not only because they contain more storylets, more outgoing links on average and cycles, but also because they include transitions that depend on the reader's ability to solve riddles. For example, in "The 88 dolmadakia" (Kalendis, 1997) we read: "Continue on the page whose first digit is found *whereto* ends and the second digit is hidden in *axis*" vs. "Continue on the page whose first two digits are twice as many as the dwarfs in the tale of Whitesnow and the third is as many as the eyes of a Cyclop". Therefore, more "ergon" is necessary so that young readers find their way, earning these books the characterization "ergodic" in Aarseth's

NHT'13, May 1, 2013, Paris, France.
Copyright 2013 ACM 978-1-4503-2005-4/13/05... $15.00

[1] Possibly moving towards "algorithmic criticism", borrowing S. Ramsay's term [21].

onomatology. On the other hand, quoting from [20]

> “Ergodic” is borrowed by Aarseth from the field of ergodic theory, where it means something else entirely. The present use of the term is justified by the etymological roots of the word, which are in the Greek words for “work” and “path”.

Over the years, the discord alluded in the above quotation has caused some unease and Aarseth to declare[2]

> If I as a game researcher have offended economists or mathematicians by appropriating ‘ergodic’, (...) I made clear the meaning of ‘ergodic’ in my book, and most readers seem to have had little trouble understanding that it was not dependent on its use in other scientific fields.”

The unease, however, continues[3]. We remark that even though the word “ergodic” is the combination of two Greek words, it did not exist in classical Greek but first appeared in German as “Ergode”’ in a 1884 landmark paper by Ludwig Boltzmann[6] to describe a stationary ensemble with only a single integral of motion. It was then translated into English and similarly in other languages, including modern Greek. What is interesting in our discussion is that the coinage of the German word in the “original” Physics context has also been the subject of some controversy, the question being whether Boltzmann derived it from ἔργον and ὁδός (as suggested by Paul and Tatiana Ehrenfest in their 1912 review of Boltzmann’s legacy) or from ἔργον and εἴδος (form, shape)[4]. We refer to [22] as well as the case presented by Gallavotti in support of the latter etymology in [8, 11]. Moreover, “Ergode”’ is phonetically close to the classical Greek word ἐργώδης that derives from the latter interpretation and means difficult, troublesome; see also the discussion by Mathieu in [17].

Based on previous findings, we argue that the use of the term is justified not only on the basis of its etymology (irrespective of which one is correct or better) but also on the basis of the use of the term in mathematics. We recall the “spectral device” that was used in [16] in order to rank storylets for books whose graph was a general graph (possibly containing cycles). Starting from an underlying matrix encoding the links between storylets that was made stochastic so that entries along each matrix column correspond to probabilities of transition between the source storylet for the column and each of the target storylets in each row, say S, storylet ranking was based on the dominant eigenvector, say p, (also called Perron eigenvector that corresponds to the largest eigenvalue that is known to be 1) of the “parametric Google matrix” $G(\mu) = \mu S + (1-\mu)H$. Matrix H is a “teleportation” matrix of the form $ve^\top$ (thus rank-1), where v is stochastic and e is the all 1’s vector, and $0 < \mu < 1$ is a “damping factor” ensuring that there is a single eigenvalue of modulus 1 and so that the eigenvector p is unique. We note that in [16] we had adjusted all graphs so that they contain a link from the ending vertex to the starting one in line with the option “If you did not like this ending, read from the beginning, this time choosing different links”, offered at all endings of the third set of CHILDIF . This made all matrices in CHILDIF irreducible. We also computed the dominant eigenvalues of the S matrices and found their maximum eigenvalue (which was 1 because of stochasticity) to be larger than all others in modulus, hence the matrices were primitive. Because of these facts, the dominant eigenvectors of S and $G(\mu)$ are strictly positive and each can be used to define a unique storylet ranking. Moreover, the use of a damping factor permits the intepretation of PageRank in terms of a “random surfer” that follows the existing link structure with probability μ or chooses a node from those in the graph with a probability distribution determined by the teleportation vector v.

Consider next the use of the term “ergodic” in Markov chains [15]. Both S and any $G(\mu)$ for each matrix from the CHILDIF collection can be considered to be a transition probability matrix for a discrete-time, finite-state, stationary Markov chain whose states correspond to the storylets. At any given time, someone reading the book will be looking at some storylet (so that the specific state would be “occupied”). In the language of Markov chains, because of irreducibility and primitivity, all states form a single ergodic set and the chain is ergodic. Moreover it holds that $p = \lim_{k\to\infty}(G(\mu))^k p_0$ for any initial stochastic vector p_0. This statement is an ergodic theorem for Markov chains. In our context, it says that the reader leafing through the pages of the book according to the transition probabilities of any $G(\mu)$, in the limit s/he could be reading any storylet of the book with non-zero probability (equal to the corresponding entry in p) irrespective of where he started from. We thus observe that Markov chains are not only suitable for analyzing ergodic books but also provide a connection of the genre with one (common) use of the term in mathematics. From this perspective, the term selected by Aarseth turns out to be suitable for reasons that go beyond etymology.

It is worth noting that Hoenkamp and Song in [13] considered the possibility of modeling a document as an ergodic Markov process based on the HAL representation (cf. [7]) and that PageRank-like methods have been used in text analysis and summarization; cf. [5, 10, 18]. To the best of our knowledge, however, our discussion is the first to provide an explicit connection between the ergodic literary genre and mathematical techniques deploying ergodic theory.

3. ENTER CONTENT

Our investigations thus far were based strictly on the storylet link structure. In [16] we concluded that link-based ranking (conducted using CHILDIF as dataset) and no information about the content, can uncover latent information about the books. This information has proven useful in the (hand) construction of CMs. An example of such a concept map can be found in Figure 6 of that paper. Given that “hypertext = content + links” (in our case “content” is the text in the storylets) a natural next step is to consider what can be done if we also include content information. Indeed, blending content with links to enhance search and clustering is an important topic in hypertext [24] and Web science; see

[2] `http://grandtextauto.org/2005/08/12/clarifying-ergodic-and-cybertext/?year=2005&monthnum=08&day=12&page`

[3] As M. Bernstein remarked in his recent blog “I need a word and it’s not ergodic”, http://www.markbernstein.org/, Jan. 22, 2013, “But Espen Aarseth used ‘ergodic text’ to mean something entirely different, and that usage is now established.”

[4] See for example the on-line Liddel & Scott Greek-English Lexicon at `http://www.tlg.uci.edu/lsj/#eid=43379&context=lsj&action=from-search`.

e.g. [12, 19, 23, 25].

A long term objective is to automate the construction of CMs. That is, we seek to provide an automatic, abstract summarization of the most important concepts from different plots encountered in any given book and represent them as graphs with links corresponding to key decisions in the plot. Our shorter term goal is to consider the role of content and how does this relate to the link-based one. To undertake the necessary text-mining tasks we deployed our Text-to-Matrix Generator (TMG version 6.0) MATLAB Toolbox [26]. The current version of TMG contains modules for the basic IR tasks of Indexing, Dimensionality Reduction, Non-Negative Matrix Factorization, Clustering, Classification and Retrieval. We also used the MATLAB Brain Connectivity Toolbox (BCT)[5].

3.1 Experiment

We illustrate our approach using the "Mystery of the Maya" book in the CYOA collection (abbreviated as `CYOA_MM`). We recall from [16] that the digraph (depicted in Fig. 1) is acyclic (thus a DAG) and consists of 113 nodes and 116 links. Moreover, there are 39 possible endings and a total of 106 plots[6].

We first applied BCT to enumerate all plots and then concatenated the corresponding storylets in order to create a single text file for each plot. We then used the TMG indexing module in order to create the *term-by-plot matrix* (tpm for short), removing common words as well as numeric and alphanumeric terms and applying stemming and `tf-idf` weighting. The size of the resulting tpm was 1578 × 106.

The next step consisted of clustering the 106 plots using spherical k-means with random initialization ([9]) as implemented in the clustering module of TMG. For this we selected $k = 6$ clusters in line with our (subjective) finding of the main concepts encountered when reading the book in preparing [16].

Depicting all clusters is not possible here because of length constraints. Instead, in Figs. 2-5 we show clusters 1, 3, 4 and 6 and mark the circumferences of the ending nodes of each (shown bold in a b/w printout and colored on screen). Except for cluster 6, all others are depicted by including also nodes that do not belong to the clusters. In reading each figure in order to recreate the plots that belong to the cluster one must follow the (typically ascending) walk from each ending node towards the "local root" (`loc_root`) that is a common ancestor of all ending nodes in the cluster (in cluster 1 this is node 12). Using this approach, we consider the walks from the starting storylet (the single source vertex, labeled "1") to the local root node for each cluster as delineating a shared "plot prefix". In `CYOA_MM` and several of the CHILDIF books, this plot prefix did not seem to distinguish the clusters.

We make the following observations: *i*) Each cluster appears to share one dominant concept as tabulated in Table 1.

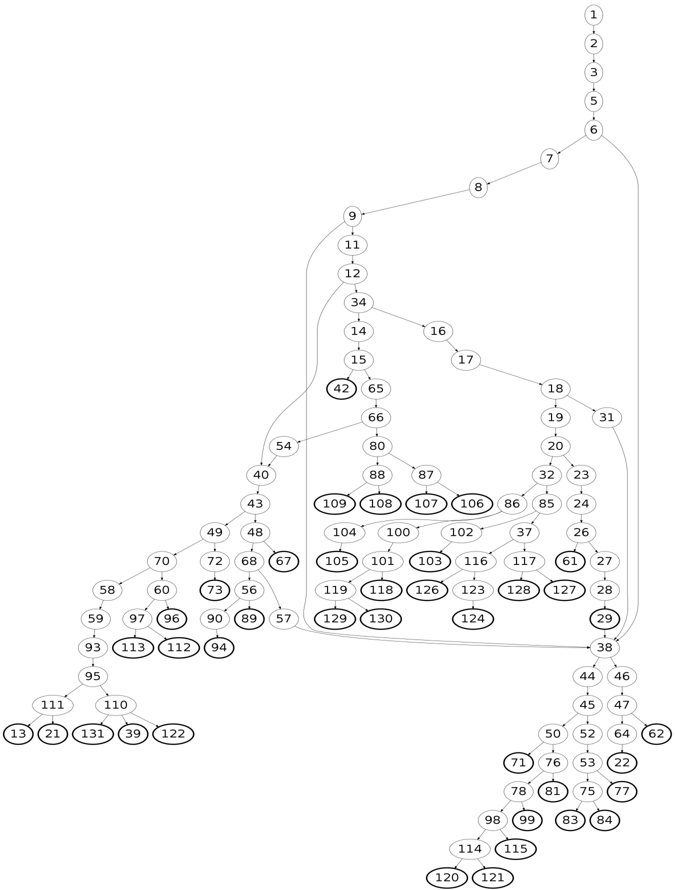

Figure 1: Graph of `CYOA_MM` (Mystery of the Mayas) from [16]. Circles in bold denote ending nodes.

cluster	Concepts	`loc_root (link)`
1	Choose to stay with the priests.	12 (12,34)
2	Choose to go to present or past.	5 (5,6)
3	Choose to explore temples and end up with the police.	44 (44,45)
4	Choose to stay in the defend team of warriors group.	43 (43,49)
5	Stories with spacecrafts.	46 (46,47)
6	Choose to explore temples and end up with the army.	52 (52,53)

Table 1: Clustering Results

We believe that this is due to a large extent to the fact that clusters shared a lot of common information because of the common plot prefix but also because of the nearby placement of their respective ending nodes. In particular, we observed that the common plot prefix is the one responsible for the assignment of each plot to the proper cluster. On the other hand, the common links starting from each local root determine the dominant concept of the cluster something that explains the importance of this node. In fact (we hope that)

[5] Toolbox authored by O. Sporns; `http://www.indiana.edu/~cortex/connectivity.html`.

[6] As in [16], by "plot" we mean a possible reading of the book woven by putting together a sequence of storylets that respects the link structure, from starting to some terminal page (cf. [4]); see also [3]. In graph theoretic terms this would be a walk from a source vertex to one labeled as ending, that makes for a "valid" reading; for example, in the presence of cycles, on could disallow mutiple repetitions of a path.

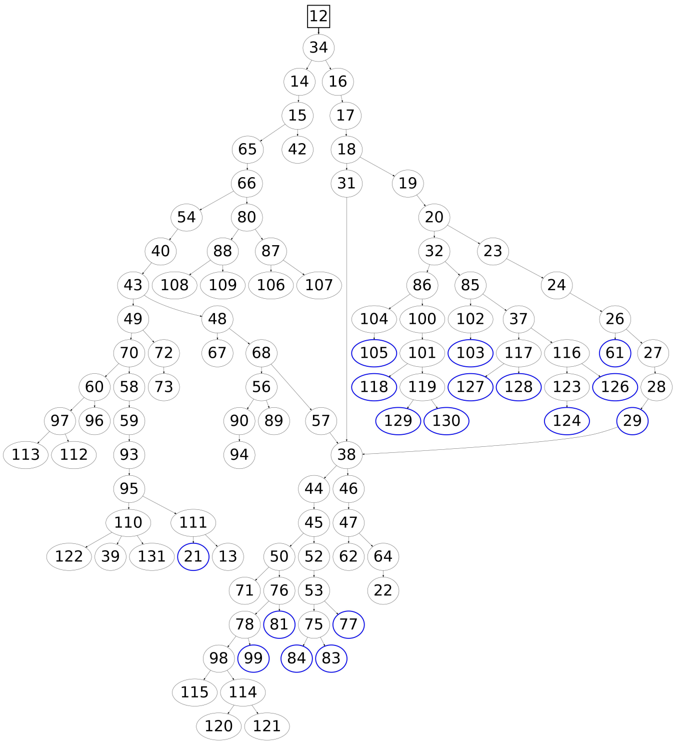

Figure 2: Subgraph with ending nodes of cluster 1

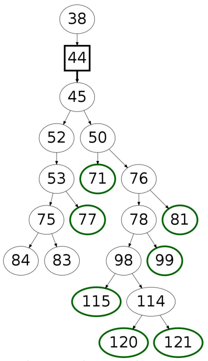

Figure 3: Subgraph with ending nodes of cluster 3

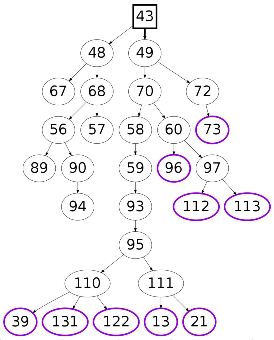

Figure 4: Subgraph with ending nodes of cluster 4

this could have been obtained by summarization and topic identification, something that remains to be done.

ii) As an illustration, consider the plots in cluster 4 (Fig. 4). These plots describe adventures in which the hero belongs to a group of Mayan warriors ("raiders" or "defenders").

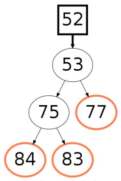

Figure 5: Subgraph with ending nodes of cluster 6

Looking at the graph and ascending the nodes starting from the darker (violet) ones that are end nodes of this cluster, one eventually encounters node 43 that connects the subgraph of cluster 4 with the remaining graph. It turns out that the choice in storylet 43 is whether the reader prefers to be a defender or a raider which is a key question for plot development. Interestingly, without the content based analysis, this information would have stayed hidden.

iii) In Fig. 6 we modified the CM obtained using only link-based information to include concepts obtained from the content-based analysis just described. We observe that the new information (dot-bordered rhombi) corresponding to the clustering based on content refines the CM by introducing concepts that are closer to ending nodes but are still of significant importance in plot development.

4. CONCLUSIONS

We argued that the characterization "ergodic" for CYOA-type books can be justified not only because of etymology but also because scientific concepts of ergodicity from the mathematics of Markov matrices are useful in analyzing this type of literature (cf. [16]). We also experimented with text mining on the content of these books and described some of our findings. As with many other applications where content and links co-exist, the question that opens up for these texts is how to best blend similarity with adjacency matrices? We hope to report on this in the future.

5. ACKNOWLEDGMENTS

We are grateful to the two reviewers for their comments. One reviewer's rightful (and playful) criticism and discussions with Aristoula Georgiadou and George Xydopoulos led us to *onomatology*, in its meaning "the science of the formation of names or terms" [1]. We thank them both.

6. REFERENCES

[1] *The New Shorter Oxford English Dictionary.* Clarendon Press, Oxford, 1993.

[2] E. Aarseth. *Ergodic Literature.* The Johns Hopkins University Press., Baltimore, MD, 1997.

[3] A. Bell. *The Possible Worlds of Hypertext Fiction.* Palgrave MacMillan, 2010.

[4] M. Bernstein. On hypertext narrative. In *Proc. 20th ACM Conf. Hypertext and Hypermedia*, HT'09, pages 5–14, New York, NY, USA, 2009. ACM.

[5] R. Blanco and C. Lioma. Graph-based term weighting for information retrieval. *Information Retrieval*, 15:54–92, 2012. 10.1007/s10791-011-9172-x.

[6] L. Boltzmann. Ueber die Eigenschaften monocyklischer und andere damit verwandter Systeme. *Crelles J.*, pages 68–94, 1884/85.

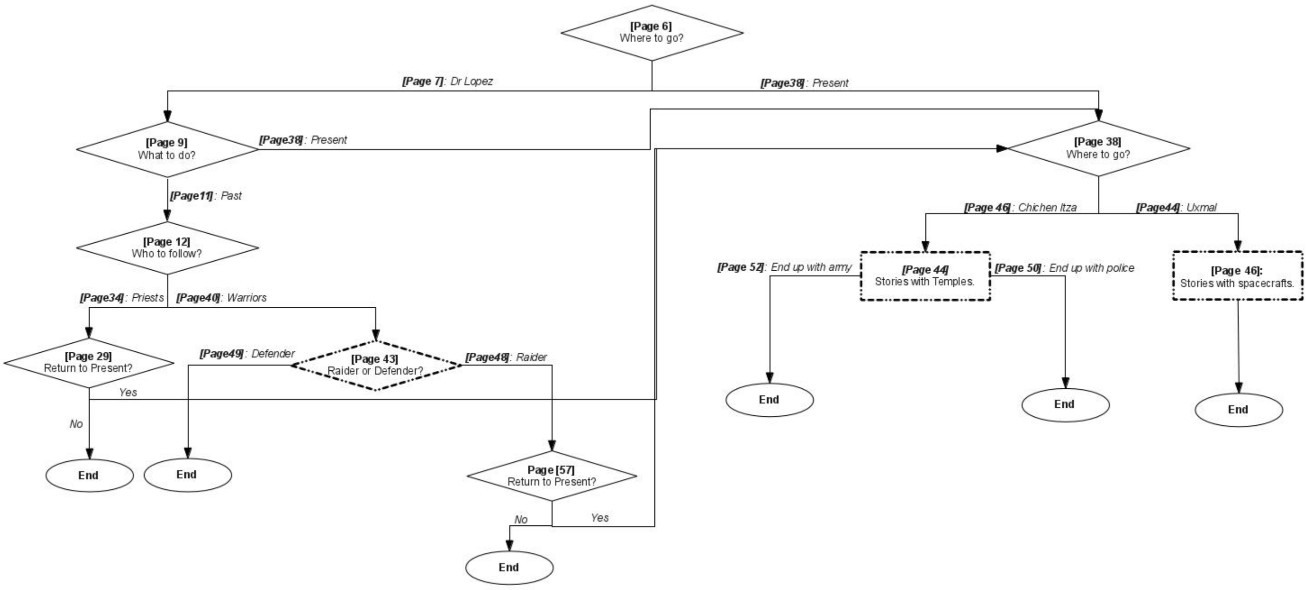

Figure 6: Concept map for `CYOA_MM` obtained from [16] (shapes with continuous line borders) modified to also take into account content-based information (shapes with dotted borders) as we described in the text. Each box represents several storylets and possible plots, but sharing the marked concept. Nodes labeled "'End" sometimes correspond to the merging of several ending nodes of the original graph.

[7] C. Burgess, K. Livesay, and K. Lund. Explorations in context space: Words, sentences, discourse. *Discourse Processes*, 25:211–257, 1998.

[8] C. Cecignani. *Ludwig Boltzmann: The Man Who Trusted Atoms.* Oxford University Press, 1998.

[9] I. S. Dhillon and D. S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1):143–175, 2001.

[10] G. Erkan and D. R. Radev. Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22:457–479, December 2004.

[11] G. Gallavotti. *Statistical Mechanics: A Short Treatise.* Springer, 1999.

[12] C. Havasi, J. Pustejovsky, R. Speer, and H. Lieberman. Digital intuition: Applying common sense using dimensionality reduction. *IEEE Intelligent Systems*, pages 24–35, 2009.

[13] E. Hoenkamp and D. Song. The document as an ergodic Markov chain. In *Proc. 27th ACM SIGIR*, pages 496–497, New York, NY, USA, 2004. ACM.

[14] N. Hoyles. *Electronic Literature: New Horizons for the Literary.* University of Notre-Dame Press, 2008.

[15] J. Kemeny and J. Snell. *Finite Markov Chains.* Van Nostrand, Princeton, NJ, 1960.

[16] E.-M. Kontopoulou, M. Predari, T. Kostakis, and E. Gallopoulos. Graph and matrix metrics to analyze ergodic literature for children. In *Proc. 23rd ACM conference on Hypertext and social media*, pages 133–142, New York, NY, USA, 2012. ACM.

[17] M. Mathieu. On the origin of the notion 'Ergodic Theory'. *Expositiones Mathematicae*, 6:373–377, 1988.

[18] R. Mihalcea and D. Radev. *Graph-based Natural Language Processing and Information Retrieval.* Cambridge University Press, 2011.

[19] D. S. Modha and W. S. Spangler. Clustering hypertext with applications to web searching. In *Proc. 11th ACM Conf. Hypertext and hypermedia*, pages 143–152, New York, NY, USA, 2000. ACM.

[20] N. Montfort. Cybertext killed the hypertext star: The hypertext murder case. *Electronic Book Rev.*, 2000.

[21] S. Ramsay. *Reading Machines: Toward an Algorithmic Criticism.* University of Illinois Press, 2011.

[22] J. Uffink. Boltzmann's work in statistical physics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Winter 2008 edition, 2008. See also http://plato.stanford.edu/entries/statphys-Boltzmann/notes.html.

[23] Y. Wang and M. Kitsuregawa. Evaluating contents-link coupled web page clustering for web search results. In *Proc. ACM CIKM*, pages 499–506, New York, NY, USA, 2002. ACM.

[24] R. Weiss, B. Vélez, and M. Sheldon. HyPursuit: a hierarchical network search engine that exploits content-link hypertext clustering. In *Proc. 7th ACM Conf. Hypertext*, pages 180–193, New York, NY, USA, 1996. ACM.

[25] S. Yan, D. Lee, and A. Wang. Costco: Robust content and structure constrained clustering of networked documents. In *Proc. 12th Int'l. Conf. Computational Linguistics and Intell. Text Proc. - Vol. Part II*, CICLing'11, pages 289–300, Berlin, Heidelberg, 2011. Springer-Verlag.

[26] D. Zeimpekis and E. Gallopoulos. TMG: A MATLAB toolbox for generating term document matrices from text collections. In J. Kogan, C. Nicholas, and M. Teboulle, editors, *Grouping Multidimensional Data: Recent Advances in Clustering*, pages 187–210. Springer, Berlin, 2006.