Relevance-driven Clustering for Visual Information Retrieval on Twitter

Mohamed Reda Bouadjenek
University of Toronto
Toronto, Ontario M5S 3G8, Canada
mrb@mie.utoronto.ca

Scott Sanner
University of Toronto
Toronto, Ontario M5S 3G8, Canada
ssanner@mie.utoronto.ca

ABSTRACT

Geo-temporal visualization of Twitter search results is a challenging task since the simultaneous display of all matching tweets would result in a saturated display. In such settings, clustering search results can assist users to scan only a few coherent groups of related tweets rather than many individual tweets. However, in practice, the use of unsupervised clustering methods such as k-means does not necessarily guarantee that the clusters themselves are relevant. Therefore, we develop a novel method of relevance-driven clustering for visual information retrieval to supply users with highly relevant clusters representing different information perspectives of their queries. We specifically propose a Visual Twitter Information Retrieval (Viz-TIR) tool which based on a fast greedy algorithm that optimizes an approximation of an expected F1-Score metric to generate these clusters. We demonstrate its effectiveness w.r.t. k-means and a baseline method that shows all top matching results on a scenario related to searching natural disasters in US-based Twitter data. Our demo shows that Viz-TIR is easy to use and more precise in extracting geo-temporally coherent clusters given search queries in comparison to k-means, thus aiding the user in visually searching and browsing social network content. Overall, we believe this work enables new opportunities for the synthesis of information retrieval as well as combined relevance and display-aware optimization techniques to support query-adaptive visual information exploration interfaces.

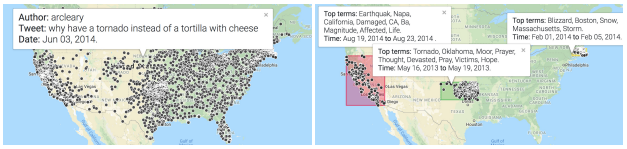
Keywords: Visual Search Interfaces; Relevance-driven Clustering.

ACM Reference format:
Mohamed Reda Bouadjenek and Scott Sanner. 2019. Relevance-driven Clustering for Visual Information Retrieval on Twitter. In *Proceedings of 2019 Conference on Human Information Interaction and Retrieval, Glasgow, United Kingdom, March 10–14, 2019 (CHIIR '19)*, 5 pages.
DOI: https://doi.org/10.1145/3295750.3298914

1 INTRODUCTION

Traditional search engines such as Google or Bing display search results in a vertical list of textual summaries. However, this display mode is certainly not adapted for search results over Twitter content, since related tweets are often geographically and temporally localized. Moreover, given the massive volume of available information in Twitter, displaying all relevant tweets for a given query prevents the visual extraction of relevant information as it results in a saturated and unreadable display [1–3]. In such settings, standard clustering methods such as k-means [4] can be used to address this issue, based on the assumption that documents in the same cluster behave similarly with respect to information needs. This is known as the *cluster hypothesis* [5–8]. While the approach of query-specific clustering has been widely explored in the information retrieval literature [9–14], all methods tend to use unsupervised clustering methods such as k-means that do not necessarily guarantee that the clusters themselves are relevant. Therefore, the development of novel visual interfaces based on relevance-driven clustering is necessary to supply the users with highly relevant clusters representing different information perspective of their queries. In this work, we propose a Visual Twitter Information Retrieval (Viz-TIR) tool for relevance-driven and display-aware clustering and presentation of Twitter search results based on timestamp, location, and text content of tweets.

To make the task of clustering in visual information retrieval more concrete, we introduce an example scenario. Consider the case of searching a multiyear Twitter corpus for content related to natural disasters. As shown in Figure 1(a), a typical visual approach would be to provide all matching tweets in an interactive visual interface. Clearly in this case, displaying all matching tweets simultaneously results in a saturated and unreadable display that will take the user a large amount of time to sift through. To ease the task of browsing search results, a clustered results display like that shown in Figure 1(b) can be used to restrict the displayed information to highly relevant clusters that cover a large fraction of



(a) Baseline display showing all results. (b) Visually clustered results display.

Figure 1: (a) An interface for visual information retrieval showing all geolocated tweets that match a query related to natural disasters in a multiyear Twitter corpus, where *all* matching tweets are shown. (b) A clustered version of the same search results showing three clusters (i.e., bounding boxes) of tweets, in this case corresponding to three well-defined natural disasters: (i) a blizzard in Boston in February 2014, (ii) a tornado in Oklahoma in May 2013, and (iii) an earthquake in California in August 2014.

CHIIR '19, March 10–14, 2019, Glasgow, United Kingdom

ACM ISBN 978-1-4503-6025-8/19/03.
DOI: https://doi.org/10.1145/3295750.3298914

relevant content. Hence it is critically important to define a clustering objective that optimizes for relevance, coverage, and visually coherent presentation of the clustered search results.

To the best of our knowledge, Viz-TIR is the first tool to address relevance-driven clustering of search results in Visual Information Displays for social media, and moreover, to do so as the direct optimization of spatial, temporal, and content-based cluster parameters w.r.t. surrogates of F1-Score to balance cluster precision and recall.

2 VizTIR CORE DESCRIPTION

In this section, we first define the mathematical notation used in this paper, then we proceed to propose Expected F1-Score (EF1) as a clustering objective well-suited to our task. Finally, we briefly describe a fast greedy relevance-driven clustering algorithm for optimizing EF1 that drives our real-time Viz-TIR interface.

2.1 Mathematical Notation

Throughout this paper we use the following mathematical notation:

- An tweet j has three types of associated metadata: (i) textual content, which is composed of a set of terms of size n, (ii) a timestamp t_e, which may represent the creation date of j, and (iii) a position coordinates (x_e, y_e).
- Three variables $I(j)$, $B(j)$ and $S(j)$ are associated with each tweet j: $I(j)$ refers to whether a tweet j is selected; $B(j)$ is a Boolean random variable indicating the relevance of a tweet j; $S(j)$ is a probabilistic score indicating the relevance of a tweet j w.r.t. a query. $B(j)$ follows a *Bernoulli* distribution with parameter $S(j)$, and hence, the expectation of $B(j)$ is $S(j)$, i.e., $\mathbb{E}_{\mathbb{S}}[B(j)] = S(j)$.
- GC is the global set of tweets with total size $|GC| = m$.
- The set of tweets selected to match a user query is labeled E with $E \subseteq GC$; we use E^* to refer to further subsets of tweets of clusters, i.e., $E^* \subseteq E$. Note that the size of E can be represented as the sum of $I(j)$ among the global collection GC. Therefore, we have $|E| = \sum_{j=1}^{m} I(j)$.
- We label the set of ground truth relevant tweets as the relevant set RS consisting of $|RS|$ elements. Note that $|RS|$ is equal to the sum of $B(j)$ among the global collection GC. Therefore, we have $|RS| = \sum_{j=1}^{m} B(j)$.

2.2 Deriving Expected F1-Score (EF1)

We adopt the Boolean relevance framework standard in information retrieval [15]. Thus, we assume that any information element j has a ground truth relevance assessment $B(j)$ available at evaluation time. Because clusters are equivalent to Boolean retrieval (they either select or do not select elements) and we have a probabilistic estimate of relevance $S(j)$, we propose to evaluate *expected* variants of standard precision, recall, and F1-score. While both precision and recall can be trivially optimized through pathological solutions (maximizing recall would select all information elements while maximizing precision would select the single highest probability information element), expected F1-score is a non-pathological objective that balances both expected precision and recall.

Recalling our previous definitions, given a set of selected information elements E and a relevant set RS, by rearranging and cancelling terms, the F1-Score of E can be expressed as follows:

$$F1(E) = \frac{2 \times \sum_{j \in E} B(j)}{|E| + |RS|} = \boxed{\frac{2 \times \sum_{j=1}^{m} B(j)I(j)}{\sum_{j=1}^{m} I(j) + \sum_{j=1}^{m} B(j)}} \tag{1}$$

Taking a 1st order Taylor expansion, we have the following expectation approximation $\mathbb{E}(X/Y) \approx \mathbb{E}(X)/\mathbb{E}(Y)$ for two dependent random variables X and Y [16]. Hence, given that $B(j)$ is a Boolean random variable, we define an *approximated expected recall* as:

$$EF1(E) = \mathbb{E}_{\mathbb{S}}\left[\square\right] \approx \frac{2 \times \sum_{j=1}^{m} \mathbb{E}_{\mathbb{S}}[B(j)]I(j)}{\sum_{j=1}^{m} I(j) + \sum_{j=1}^{m} \mathbb{E}_{\mathbb{S}}[B(j)]} = \frac{2 \times \sum_{j=1}^{m} S(j)I(j)}{\sum_{j=1}^{m} I(j) + \sum_{j=1}^{m} S(j)} \tag{2}$$

2.3 Greedy relevance-driven clustering

Now we greedily optimize $EF1(E)$. In this work, we assume that three types of clustering "parameters" are used to optimize clusters: Keywords, Time, and Space. A cluster is generated by *conjoining* these three selection parameters. In the following, we describe how to greedily optimize each of these selection parameters by iterative pruning, how to combine them for producing relevance-driven clusters, and finally, how to generate multiple relevant clusters.

2.3.1 Greedy Keyword Selection algorithm. Given a set of information elements matching a user query, the Greedy Keyword Selection algorithm aims to select a set of keywords in order to exclude a subset of tweets containing these keywords for the purpose of maximizing the EF1-Score. Formally, the algorithm aims to select an optimal subset of k terms $T_k^* \subset T_E$ (where $|T_k^*| = k$ and T_E are terms of the initial set of elements) to exclude tweets containing these keywords for optimizing the EF1-score. This is achieved by building T_k^* in a greedy manner by choosing the next optimal term t_k^* given the previous set of optimal term selections $T_{k-1}^* = \{t_1^*, \ldots, t_{k-1}^*\}$ (with $T_0^* = \emptyset$) using the following criterion:

$$t_k^* = \underset{t_k \notin T_{k-1}^*}{\arg\max} [EF1(E^* \text{ that don't contain } \{t_1^*, \ldots t_k^*\})] \tag{3}$$

where E^* is a subset of the initial tweet set E that don't contain the keywords $\{t_1^*, \ldots t_k^*\}$.

2.3.2 Greedy Time Selection algorithm. The idea behind the time-based greedy selection algorithm is as simple as finding a time window range $[t_{start}, t_{end}]$ for *temporal coherency* of clusters, which allows to select a subset of tweets $E^* \subseteq E$ falling in that time window, with E^* having the highest EF1-Score. Formally, given a list of elements $E = \{j_{t_1} \leq \cdots \leq j_{t_n}\}$, where "$\leq$" specifies the timestamp order, we propose to use *binary partitioning search* (BPS) to find the best set that optimizes EF1.

2.3.3 Greedy Spatial Selection algorithm. It is critical to main *visual coherency* of clusters and bounding boxes are one way to visually bound regions that facilitate direct optimization. The aim of the spatial greedy selection algorithm is to return coordinates $[(x_{min}, y_{min}), (x_{max}, y_{max})]$ representing the EF1-Score maximizing bounding box represented by the lower and upper bound coordinates – respectively (x_{min}, y_{min}) and (x_{max}, y_{max}). This 2D

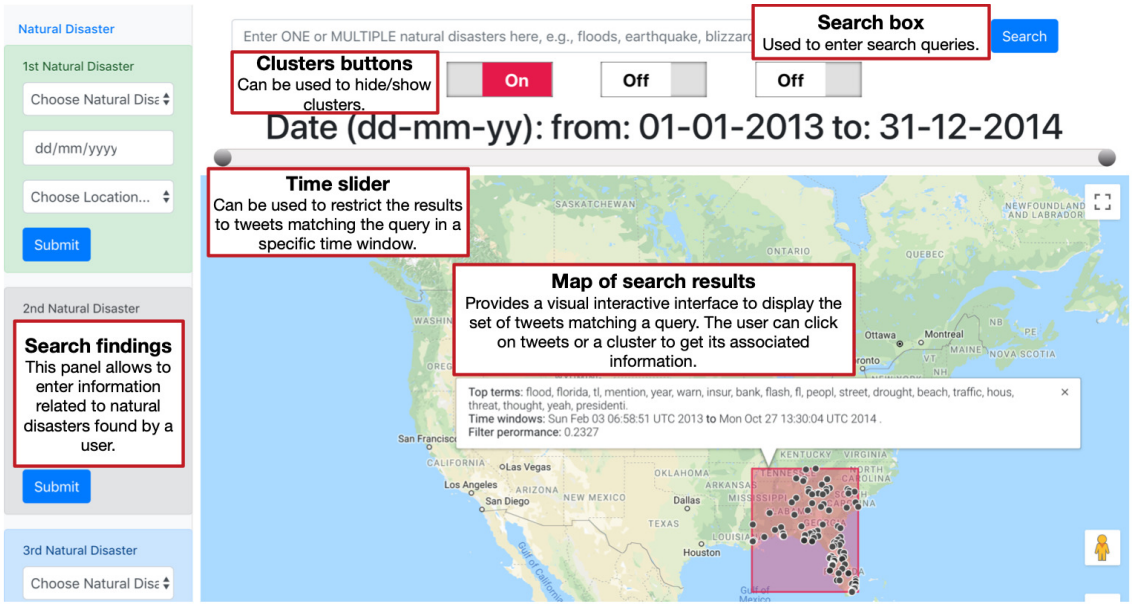


Figure 2: Main interface of the demonstration.

problem is similar to the previous one dimensional problem of finding the best time window. Therefore, we propose to sequentially apply a BPS on the x-axis to determine (x_{min}, x_{max}), then on the y-axis to determine (y_{min}, y_{max}).

2.3.4 Relevance-driven clustering algorithm. To obtain a cluster combining the above selection parameters, we propose a greedy algorithm, which at each iteration applies all the above selection algorithms and chooses the one that improves the most EF1. The selected cluster is updated with its new setting and the iteration continues. Iterations terminate when no selection algorithm can unilaterally improve EF1 and the final cluster is returned.

2.3.5 Multiple Cluster Selection Wrapper. In practice a single cluster chosen by the previously described algorithm will narrow the user in a single "information perspective". However, there will likely be multiple perspectives and so the user should have a choice of multiple clusters. Consider Figure 1: this actually shows three different spatial bounding boxes corresponding to three different events provided by three clusters. Here, we provide a greedy approach for providing a ranked list of clusters. The algorithm itself is quite simple: after the first cluster is produced, all selected tweets in that cluster have their scores $S(j)$ zeroed out. The relevance-driven clustering algorithm is then run again, where it will inherently focus on a different content set.

3 DEMONSTRATION OVERVIEW

The demonstration will be illustrated using a scenario related to finding natural disasters discussed in a collection of tweets. We used Twitter data crawled using the Twitter Streaming API for two years spanning 2013 and 2014 [17] with the following restrictions intended for user experimentation: (i) the dataset was restricted to users located within the US, (ii) non-English tweets were filtered out, (iii) only the tweets related to 12 natural disasters were kept – tweets related to other natural disasters were removed. These natural disasters are temporally, and geographically disjoint – a storm, a hurricane, a drought, two floods, two earthquakes, two tornadoes, and three blizzards. Finally, (iv) false positive tweets mentioning natural disaster keywords but not related to a particular natural disaster were intentionally included. The final dataset contains 39,486 tweets with 5,075 relevant natural disaster tweets.

As shown in Figure 2, the demonstration's main user interface allows users to enter a search query with search results then shown on an interactive map used to browse the results. The user can interact with the map by panning and zooming and also by clicking on tweets and clusters to view their content (for clusters, we display a summary in terms of selected keywords). Also, the user will be able to use a time slider bar to restrict the results to tweets in a specific time window.

In the scenario of this demonstration, the user will experience finding three different natural disasters using three different algorithms. In addition to Viz-TIR, two different algorithms are used to show the effectiveness of Viz-TIR, including: (i) a baseline method which displays all tweets that match a query (see Figure 3) and (ii) X-Means [18] (an extension of k-means, which tries to automatically determine the number of clusters) – a baseline method for clustering the top relevant results (see Figure 5). At each step, the user will be asked to enter information related to each natural disaster they may identify, including the type of the natural disaster, its location (US state), and the date on which they think the disaster first occurred.

In this demonstration, a user can perform interactive searches such as the following example search scenarios:

Baseline scenario: A user may search for information on natural disaster events. Using the query box, the user enters the keyword

"earthquake". As shown in Figure 3, the baseline algorithm shows all tweets that match the query on the map using circles with a color range corresponding to the probability of relevance.

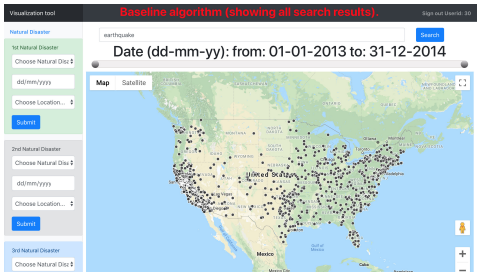


Figure 3: All results that match the query.

As shown in Figure 4, the user can restrict the list of tweets shown to a specific time window with the time slider; here a cluster of tweets appears indicating an earthquake in California during August 2014. The user may click on tweets to learn more about that event.

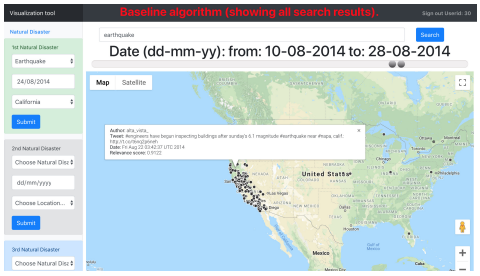


Figure 4: Sub-set of the results that match the query.

Clustering scenarios: The scenario using *k*-means or EF1 is similar. For example, the user may search multiple natural disasters using the multi-term query "earthquake, blizzard, tornado". As shown in Figures 5 and 6, clusters will appear for both algorithms, which show different clustered information perspectives. The user can then click on a cluster to get a summary description.

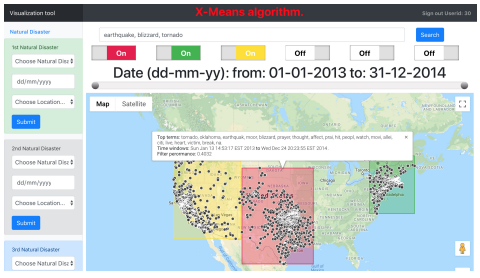


Figure 5: Clustered results using X-Means.

User Study: To determine whether our proposed relevance-driven EF1 optimization approach to clustering improves human search

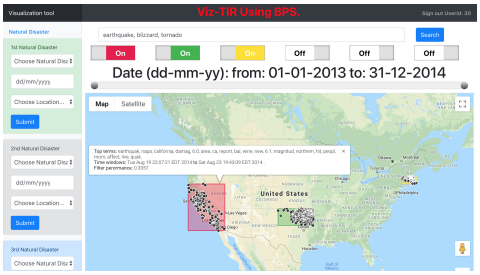


Figure 6: Compact clustered results using EF1 (Viz-TIR).

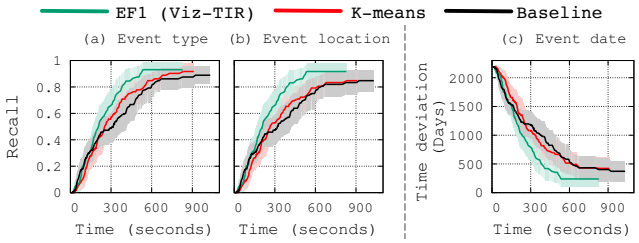


Figure 7: The performance of the algorithms measured using cumulative recall for the type and location of the natural disasters and the absolute error in the dates of the disasters.

task performance in the Viz-TIR visual search interface (in comparison to *k*-means clustering and the non-clustering baseline), we performed an evaluation of 24 users who were given the task of searching for natural disasters in the previously described data. An analysis of their performance is provided in Figure 7, which shows that on average, users have higher recall on natural disaster types and locations as well as lower error estimating the time of the disaster using EF1 clustering in comparison to other methods.

4 SUMMARY AND FUTURE WORK

In this paper, we have developed Viz-TIR, a tool to visually search Twitter content through spatial, temporal, and content-based clustering. Viz-TIR formulates clustering as a relevance-driven optimization problem w.r.t. a user-provided query. In particular, Viz-TIR leverages a fast greedy optimization algorithm to maximize an approximation of the expected F1-Score metric to generate multiple clusters for visual display. We provide a demonstration use case that compares Viz-TIR with two baselines over 2 years of Twitter content for finding information related to natural disasters.

Important areas of future work include consideration of the role of (pseudo-)relevance and other explicit or implicit feedback methods to create a tighter and more responsive user interaction loop. Furthermore, in combination with user studies and consideration of human factors, future work should also consider novel application-specific objectives, e.g., in specific visualization frameworks or based on a ranking theory of results presentation (e.g., using size or color for visual ranking emphasis).

Acknowledgement: We would like to thank Yihao Du for running the user study.

REFERENCES

[1] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J.J. van Wijk, J.-D. Fekete, and D.W. Fellner. Visual analysis of large graphs: State-of-the-art and future research challenges. *Computer Graphics Forum*, 30(6):1719–1749, 2011.

[2] Shixia Liu, Weiwei Cui, Yingcai Wu, and Mengchen Liu. A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393, Dec 2014.

[3] Guo-Dao Sun, Ying-Cai Wu, Rong-Hua Liang, and Shi-Xia Liu. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology*, 28(5):852–867, Sep 2013.

[4] Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[5] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[6] N. Jardine and C.J. van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7(5):217 – 240, 1971.

[7] Ellen M. Voorhees. The cluster hypothesis revisited. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '85, pages 188–196, New York, NY, USA, 1985. ACM.

[8] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[9] Oren Kurland. Re-ranking search results using language models of query-specific clusters. *Inf. Retr.*, 12(4):437–460, August 2009.

[10] Oren Kurland and Eyal Krikon. The opposite of smoothing: a language model approach to ranking query-specific document clusters. *Journal of Artificial Intelligence Research*, 41:367–395, 2011.

[11] Or Levi, Ido Guy, Fiana Raiber, and Oren Kurland. Selective cluster presentation on the search results page. *ACM Trans. Inf. Syst.*, 36(3):28:1–28:42, February 2018.

[12] Fiana Raiber and Oren Kurland. Exploring the cluster hypothesis, and cluster-based retrieval, over the web. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2507–2510, New York, NY, USA, 2012. ACM.

[13] Ismail Sengor Altingovde, Rifat Ozcan, Huseyin Cagdas Ocalan, Fazli Can, and Özgür Ulusoy. Large-scale cluster-based retrieval experiments on turkish texts. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 891–892, New York, NY, USA, 2007. ACM.

[14] Xiaoyong Liu and W. Bruce Croft. Cluster-based retrieval using language models. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 186–193, New York, NY, USA, 2004. ACM.

[15] Ricardo A Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 2 edition, 2010.

[16] G.M.P. van Kempen and L.J. van Vliet. Mean and variance of ratio estimators used in fluorescence ratio imaging. *Cytometry*, 39(4):300–305, 2000.

[17] Zahra Iman, Scott Sanner, Mohamed Reda Bouadjenek, and Lexing Xie. A longitudinal study of topic classification on twitter. In *ICWSM*, pages 552–555, 2017.

[18] Dan Pelleg and Andrew W. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 727–734, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.