An Approach for Discovery of Complex Events and Hierarchies

Rashidah Namisanvu5, Makerere University, Uganda and
Benjamin Kanagwa, Makerere University, Uganda

Abstract

This paper presents techniques for discovery of event hierarchies in event streams. Event discovery is about recognition of low level events; their relationships and how they combine to cause a composite event. The challenge is that the occurrence time of a composite event, the identity of the low level events, the number of the low level events, and relationship are not known in advance. We start by identifying a set of 'candidate events' that lead to a given composite event. We then filter the candidate events to establish the actual events that lead to the composite event. Then a causal relationship between the filtered low level events is discovered. The causal relation among the events allows generation of a hierarchical structure that shows the composition structure between low level events and the composite event. We rely on domain experts and literature to identify the initial set of low level candidate events. To filter candidate events, we use a historical event stream. We develop an approach based on heuristics and similarity measures to identify the structural relations between low level events and composite event. The discovered structure of the events is then validated using domain experts. The approach was developed using a case study of financial crisis with the historical news corpus archived by major news networks, particularly CNN as the event stream.

Categories and Subject Descriptors: D.2.7 [**Software Engineering**]: — General Terms: Design

Additional Key Words and Phrases: Causality, Event discovery, Hierarchies

IJCIR Reference Format:

Rashidah Namisanvu and Benjamin Kanagwa Kizza, Joseph. An Approach for Discovery of Complex Events and Hierarchies. Vol. 9, Issue.2 pp 55 - 66. http://ijcir.mak.ac.ug/volume9-issue2/article5.pdf

5 Author's Address: Rashidah Namisanvu and Benjamin Kanagwa,, Makerere University, Uganda

International Journal of Computing and ICT Research, ISSN 1818-1139 (Print), ISSN 1996-1065 (Online), Vol.9, Issue 2, pp. 54 - 64, December 2015.

1. INTRODUCTION

Complex Event Processing (CEP) is a set of technologies [Luckham 2003; Margara and Cugola 2011] that allow incremental processing of information as it arrives in order to identify high level situations of interest or composite events, starting from low level primitive event notifications. Composite events are specified through user-defined queries, or rules, which express how to select, manipulate, and combine primitive events. The rules define composite events starting from patterns of primitive ones, involving content- based and temporal constraints. CEP is an importance technology as it allows just-in-time detection of undesirable situations or seize opportunities as they arise.

Complex event processing [Luckham 2003] has been positioned as a means to detect known event patterns of low level events. However, in some cases the low level events that lead to high-level composite events are not known in advance. At the same time, the low level events occur with little or no context that directly relates them to the composite events they influence. The challenge is that to define rules and queries for a given composite event, the underlying low level events must be known in advance. Even in situations where the composite event can be detected as a whole by observing its effects or symptoms, a better understanding of the composite event can greatly be enhanced by studying the underlying causes and events. However this is inhibited by the missing knowledge about the underlying low level events and their relationships.

In a typical environment, there are many events generated both externally and internally from different layers of the organisation. A specific event occurrence is associated with a small set of the total events that are actually occurring and being detected. The low level events bubble through the different layers and combine in different ways to bring about the larger event. When many events occur from different sources, it is hard to make sense out of such individual occurrences [Luckham 2003]. One common approach is the use of event abstractions [Cuny et al. 1993; Kunz 1993a; 1993b; Luckham 2003] as a means of simplifying complexity and understanding of different appli- cations. Event abstractions group sets of events into higher level events [Kunz 1993a]. Displaying information at various levels of abstraction enhances the subjects' ability to diagnose various types of events [Vicente 1990]. These abstractions have revealed many problems that could limit the operations of a given organisation. To take advantage of abstractions there is need for a systematic approach to detect or discover relevant abstractions both as primitive and/or as hierarchies that may be useful in understanding the given composite event. To predict or understand the occurrence of such situations, there is need to monitor low level events. However, the smaller events do not occur in isolation but happen alongside other irrelevant events or remotely related events. The collection of all events is called the global *event cloud* [Luckham 2003]. The events relevant to a given main event is a subset of the *event cloud*. For instance, the occurrence of the economic crisis [Driscoll et al. 2003; Joh 2003] is reported to have been a combination of events such as poor governance systems, weak legal environments, pressure from treasury departments, austerity imposed on governments in order to receive aid, among others.

The underlying question is :"Given the occurrence of a main event, what are the underlying low level events? and what is the relationship between the low level events?" Figure 1 is an illustrations of the steps involved in event discovery. Starting from the event cloud, the first task is to identify the possible candidate events that relate to a specific main event of interest. The candidate events are then filtered to generate the set of actual sub events for a given event. The last task is to determine any causal dependences among the sub events. The causal dependences are then used to construct an event hierarchy.

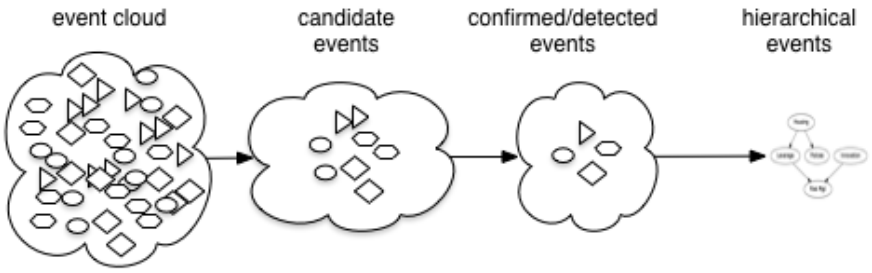


Fig. 1. Steps in Discovery of Event Hierarchies

Our approach was developed using a case study of financial crisis [Driscoll et al. 2003; Joh 2003]. However because of lack of availability of datasets, we rely on the news and articles published about financial crisis. It has been noted that more articles are published about a given event around the peak of its occurrence. The unstructured nature of the news articles, allows our techniques to be applied to many scenarios where events pass unnoticed in unstructured documents such as corporate emails, minutes of meetings, memos, phone logs and many others. For this case study, we sample more than 80000 news articles from 2001 to 2011 and by application of similarity techniques, we are able to filter the most relevant low level events as well as detect the occurrence time of the low level events. The rest of the paper is organised as follows: related work is discussed in Section 2 while Section 4 describes the process of event identification. In Section 5, we present the event detection approach and we show how to build causal relationships in Section 6. In Section 7 we validate the model followed by a conclusion in Section 8.

2. RELATEDWORK

The most closely related work is that on Detecting and Tracking of News Events[Allan et al. 1998; Nallapati et al. 2004; Radev et al. 2005; Yang et al. 1999]. In [Allan et al. 1998] a total of 15,863 chronologically ordered stories spanning 1st July 1994 to 30 June 1995. Half of the stories were randomly sampled from Reuters articles; the rest from CNN broadcasts that were manually transcribed by the Journal of Graphics Institute. In their approach they started with a total of 25 manually identified events under the TDT1 corpus. Their main focus was to detect which news stories related to one of the 25 events. Two tasks in event detection are identified. *Retrospective detection* that discovers previously unidentified events in chronologically ordered documents. *On- line detection* which identifies the onset of new events from live news feeds. The approach used is based on the conventional Information Retrieval approaches, particularly vector-space model [Soucy and Mineau 2005] for Retrospective detection and k-nearest neighbour (kNN)[Cunningham and Delany 2007] for on-line event detection. The work stops at detection of individual events and does not aim at determining causal dependences among the detected events.

In [Nallapati et al. 2004], approaches for clustering stories into events and constructing dependencies among them were suggested. They developed a time-decay based clustering approach that takes advantage of temporal localization of news stories on the same event and showed that it performs significantly better than the baseline approach based on cosine similarity. Some events however, affect the entire globe and thus attaching location to them does not help much. For instance, the financial crisis of 2008/2009 affected the entire globe and is generally referred to as the global financial crisis. Events that caused the crisis came from different places [Hellwig 2008; Mitton 2002; Obstfeld et al. 2009] but affected other places as well making location irrelevant in this case. There is a need to define events with features other than their locations.

Li et al. [2005] proposes a probabilistic model that incorporates both time and content in a unified framework. This model gives new representations of both news articles and events. Furthermore, based on this approach, an interactive RED system, HISCOVERY, which provides additional func- tions to present events is built. Whereas they provide an algorithm to represent news articles and their events, they do not discover the sub-events that relate to a particular event.

3.CASE STUDY

The financial domain, in particular the 2008/2009 financial crisis [Crotty 2009], was used as the case study for the research. Here we highlight key literature on financial crisis as part of our case study. The term financial crisis is used in a wide variety of contexts to refer to a situation where, for some reason or other, an institution or institutions lose a huge part of their value. Financial crises are a common occurrence in the world today especially in specific sectors of the economy. A financial crisis can hit a single sector of an economy and not necessarily affect the other sectors.

The causes of a financial crisis vary with the type of crisis. Although many economists have come up with causes of financial crises, there is hardly a consensus between economists on these causes. This is partly because the different perspectives of economics sometimes rival each other, and partly because perhaps every financial crisis is peculiar to itself. Figure 2 shows the different types of crises that have been occurring.

Fig. 2. Timeline of different financial crises (Adopted from [search])

According to IMF [Claessens and Kose 2013], a financial crisis is often associated with one or more of the following phenomenon: substantial changes in credit volume and asset prices; severe disruptions in financial intermediation and the supply of external financing to various actors in the economy; large scale balance sheet problems; and large scale government support. As such, financial crises are typically multidimensional events and can be hard to characterize using a single indicator.

The financial crisis of 2008/2009 is the most recent and has spread throughout the world as shown in figure 3. This crisis originated in the US and spread to the different parts of the world primarily through declines in trade [Crotty 2009]

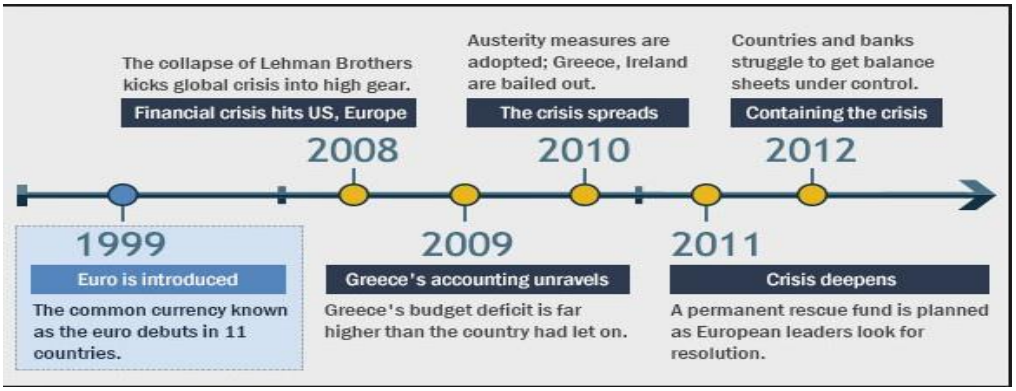


Fig. 3. Timeline of 2008/2009 financial crisis [Adopted from [Steiner 2013]]

As a case study, the problem can be re-formulated as follows:

Given the occurrence of the 2008/9 financial crisis, what are the underlying sub-events the led to its occurrence? What is the ordering and relationship among these sub-events?

4.EVENT IDENTIFICATION

To discover the underlying events of a given composite event, we needed to start with an initial set. We call this set the 'candidate events'. The process of assigning members to this set is called event identification. This process will vary for different domains. Here we used domain experts and literature review to formulate the initial set.

Working with our case study, events in table I were identified as the leading causes of the 2008/9 financial crisis. The results in the table are based on the review of the existing literature on the causes of the financial crisis and the results of the questionnaire that was used in the survey with domain experts. The events here represent a multitude of events that led to the financial crisis. The related events were grouped under one event.

Table I. Table showing the queries and their keywords

Event	**Keywords**
Housing Prices	bad loans, fore closure, mort- gage, sub-prime loan,
Risk management fail-	investment, risks, regulations,
Financial innovation	risks, product invention, spec-
Government policies	regulations, banks, deregula-
Over-leverage	borrowing, investment, specu-

5. EVENT DETECTION

Event detection is about observing an event to assign a timestamp. A financial crisis and associated low-level events happen over time, they have the start time and end time. To determine the occurrence time of the event, we used the news database. In the case of the 2008/9 financial crisis, for each news article we searched for a set of keys that relate to the event description. The search keys were extended to include synonyms for the event description. Table I shows the event description alongside associated synonyms.

The detection of an event was based on the premise that the occurrence of an event is always associated with a burst of features [Fung et al. 2007] where some features appear frequently when the event emerges and their frequencies drop when the event fades away. It was thus assumed that bouts of articles will be concentrated around the occurrence time of the event. Considering that financial information is normally given quarterly, that is in blocks of three months, the occurrence of an event was modeled over an interval of three months around the period where there is burst of news articles about an event.

Therefore an event e is defined over a time interval $t = [t0, t1]$, where t is the interval over which that event happened, $t0$ is the start time of the event and $t1$ is the end time of the event. The functions $t0(e)$ and $t1(e)$ denote the start time and end time of event e respectively. A plot of the

percentage of documents against the publication date was used to extract the values of $t0$ and $t1$. The graph generated for housing crisis is presented in Figure 4.

From Figure 4, and the other figures generated for the other events, the timestamps are sum- marised in Table II.

6. DETERMINING CAUSAL RELATIONSAND HIERARCHY

To establish the dependency or causality among the events, we used an event model similar to [Nallapati et al. 2004] as $M = (e, E)$ to be a tuple of the set of events and a set of dependencies.

Fig. 4. Graph on housing prices

Table II. The timestamps of the events as read from the graphs

Event	Timestamp
Housing crisis	Dec 2007 - Feb 2008
Risk management	Dec 2009 - Feb 2010
Financial innovation	Dec 2008 - Feb

M can be seen as a directed graph with an edge on the graph if $(e_u, e_v) \varepsilon E$. While the existence of an edge itself represents relatedness of two events, the direction could imply causality or temporal-ordering. Causal dependency means that the occurrence of event e_v is related to and is a consequence of the occurrence of event e_u.

To establish the relation between any pair of events e_u and e_v, we applied a specialised similarity technique between the keywords of e_u and the documents of e_v . For ease of exposition, we define the following:-

—$Syno(e_u)$: the set of all keywords that are synonymous with event e_u.

—$D = \{d_1, \cdots, d_n\}$: the set of all news articles.

—$Doc(e_u) = (Syno(e_u), D)$: the set of documents obtained after querying the set D with the the list of synonyms $Syno(e_u)$. $Sim(Syno(e_u), D)$ is a similarity function that takes the average of Jaccard and Cosine similarity techniques and returns a value between 0 and 1 which helps to filter the related documents from those that are not related to the query.

—$Occ(e_u, e_v) = (Syno(e_u), Doc(e_v))$: the set of the total number of occurrences of the synomyms of event e_u in the documents of event e_v

We define a causality function

$$Casual(e_u, e_v) = \frac{Occ(e_u, e_v) + Occ(e_v, e_u)}{\sum_{u=1, v=1} Occ(e_u, e_v)} \quad (1)$$

where $u \neq v$.

The definition function $Casual(e_u, e_v)$ was based on the premise that if events e_u and e_v are related, then they should appear together in several news articles. In principle, we were looking at how often the documents of e_v reference the events related to e_u. That is, the keywords of one event were used to query the documents of another event. This ranking provides a measure of causality between two events. For instance, taking the housing crisis event, the keywords are mortgages, bad loans, foreclosure, housing price and subprime loan. The keywords of housing, for instance, were applied in the set of documents for the financial innovations event and vice versa.

To establish an edge between two events e_u and e_v, we looked at the number of occurrences that are common to both and used it to gauge the level of dependency of the events.

Formally,

$$Causal(e_u, e_v) > T \cap t0(e_v) < t1(e_u) = E \quad (2)$$

where T is threshold value.

The values for $Causal(e_u, e_v)$ are summarised in the table III presented as percentages. Using

Table III. Summary of the values for $Causal(e_u, e_v)$

Causal(eu, ev	Value
(H,I), (I,H)	8.7
(H,L), (L,H)	8.0
(H,R), (R,H)	7.9
(H,P), (P,H)	8.1
(R,L), (L,R)	21.9
(R,P), (P,R)	6.7
(R,I), (I,R)	7.5
(I,P), (P,I)	7.1
(I,L), (L,I)	7.2
(P,L), (L,P)	8.5

the values obtained in table III, the values ranging from 7 - 8 were used to determine the threshold. Thus the threshold T was varied through the values 7, 7.5, and 8 to check which threshold value gives the best event model in relation to the one obtained from the experts. Taking the threshold $T = 7$, $T = 7.5$ and $T = 8$ gave the edges E shown in figure 6, figure 6, and figure6 respectively.

7. MODEL VALIDATION

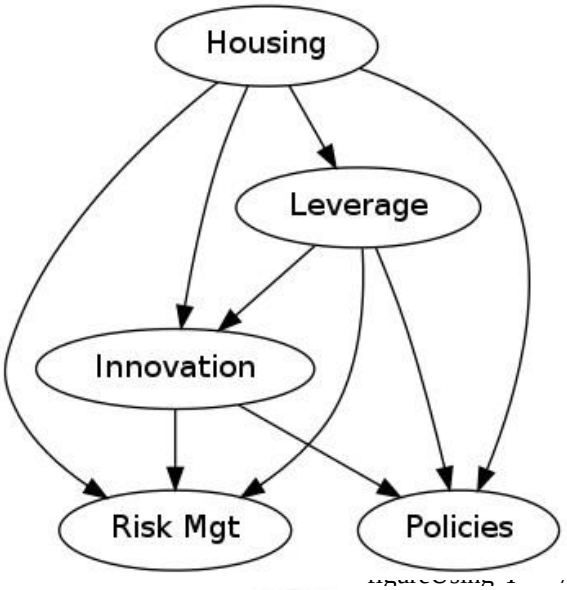
The model and the value of the threshold T were validated based on the results of domain experts. Each expert was asked to state any relationship between any given pair of events.

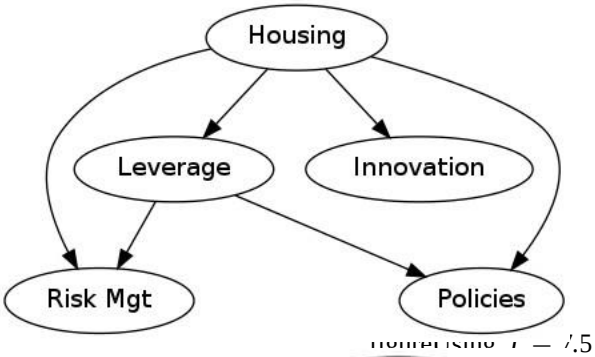
An edge was considered if 50% or more of the experts agree on the relationship. The experts event model in Figure 6 is generated from the experts' responses, following the criteria below:-

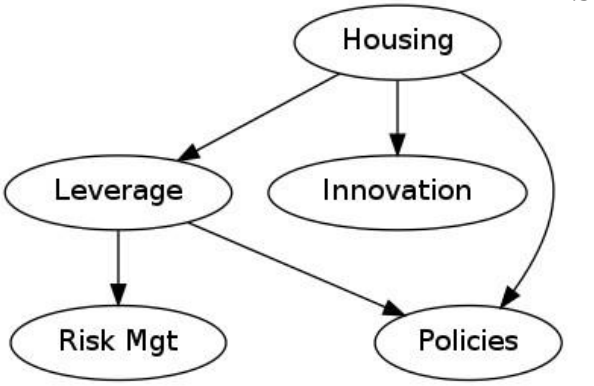
—The number of experts that gave a link between any two events was counted. In the table, the links are paired with their duals, that is, HI with IH, PR with RP as shown.

—An event with expert responses of 50% or more was taken to be a valid link. For instance, (H,I),(I,H) has a count of 30% while (I,P),(P,I) has a count of 60% making (I,P)(P,I) a valid link and (H,I),(I,H) invalid.

—The direction of the arrows was obtained from the questionnaires. For instance, given that (I,P),(P,I) is a valid link with a count of 60%, we noted that 66% of the 60% experts identified the link (I,P) and only 34% identified the link (P,I) thus giving the edge (I,P). The same approach was used for all the event pairs and Figure 6 was obtained.







figureUsing $T = 8.0$

Fig. 5. Graphs obtained for the edges E^t using the different threshold values

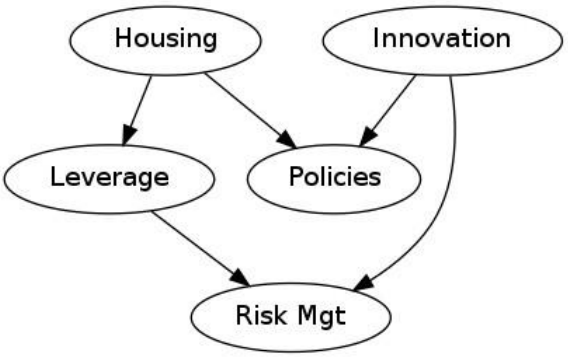


Fig. 6. Relationship between the events (Experts view)

We took the experts view as represented in Figure 6 to be the ideal model and used it validate those generated from our approach.

7.1 Determining the threshold *T*

The graphs in Figure 5 were obtained using different values of T in equation 2. We compared each of these graphs with that of the experts in Figure 6. We applied graph matching techniques from [West et al. 2001] to compute the relational distance between any two graphs as

$$MD(M1, M2) = \frac{minError(f)}{\text{sum of edges in E1 and E2}} \quad (3)$$

On applying equation 3 the graph of $T = 8$ had the least distance thus was the most similar to the experts graph. This means, the graph of $T = 8$, being that it gave the least error was the most similar to the experts' graph making our threshold $T = 8$.

7.2 Analysis of the results

We obtained a threshold of $T = 8$ which means that the events are more related to each other if they have more documents in common. That is, if the events share many documents, then most likely, one event caused the other and vise vasa.

With this threshold $T = 8$, the graph obtained was as shown in figure 6. This figure gives us a hierarchical structure of the events. According to the graph, the increase in housing prices was brought about by the financial innovations, over-leverages and the poor government policies. The over-leveraging also depended on the government policies and the failure to manage risks.

This means that at the lowest level, we have the government policies and the failure to manage risks as the leading causes of the financial crisis. The government policies were followed by financial innovations, over-leveraging, failure to manage risk and housing prices respectively. This means that those events followed in that order which does not differ much from the one presented in Figure 6.

The 2008/9 financial crisis, also referred to as the global crisis, affected many parts of the world with its root being in the United States of America. Its ability to spread that widely is an indicator that many coutries have poor or weak policies that govern the flow of their finances. These policies affect the interest rates that banks use, the way government reserves are mangaged, to mention but a few.

8. CONCLUSION AND FUTURE WORK

Given an event cloud, our approach identifies the events that relate to the goal event or the set objective and derives hierachies that convey useful information using the event keywords, and their timestamps. This approach was compared to the results obtained from the experts and proved that it can generate results that are relatively accurate. It can therefore be used in many context to detect events. Hierarchical structuring of events helps in detection of emerging events and root- cause analysis. Under root-cause analysis, a user is able to understand a given occurrence by exploring the sub-events that led to its occurrence.

More could be done to improve the proposed approach or make it available for use to a larger audience. A graphical user interface could be developed, where users can put queries and the system automatically presents the required answers in terms of event hierarchy and abstraction for a given main event. It is also possible to extend the approach to support other data sources that are not internet based. These may include structured data sources or event streams.

REFERENC
S

[1.] Allan, J., Carbonell, J. G., Doddington, G., Yamron, J., and Yang, Y. 1998. Topic detection and tracking pilot study final report.

[2.] Blundell-Wignall, A., Atkinson, P. E., and Lee, S. H. 2008. *The current financial crisis: Causes and policy issues*. OECD.

[3.] Claessens, S. and Kose, M. A. 2013. *Financial Crises Explanations, Types, and Implications*. International

[4.] Monetary Fund.

[5.] Crotty, J. 2009. Structural causes of the global financial crisis: a critical assessment of the 'new financial architec- ture'. *Cambridge Journal of Economics 33*, 4, 563–580.

[6.] Cunningham, P. and Delany, S. J. 2007. k-nearest neighbour classifiers. *Mult Classif Syst*, 1–17.

[7.] Cuny, J., Forman, G., Hough, A., Kundu, J., Lin, C., Snyder, L., and Stemple, D. 1993. The ariadne debugger: scalable application of event-based abstraction. In *Proceedings of the 1993 ACM/ONR workshop on Parallel and distributed debugging*. PADD '93. ACM, New York, NY, USA, 85–95.

[8.] Driscoll, W., Clark, J., and Association, I. D. E. 2003. *Globalization and the Poor: Exploitation Or Equalizer?*

[9.] Idea Sourcebooks in Contemporary Controversies. International Debate Education Association. Fung, G. P. C., Yu, J. X., Liu, H., and Yu, P. S. 2007. Time-dependent event hierarchy construction. In

[10.] *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM,300–309.

[11.] Hellwig, M. 2008. The causes of the financial crisis. In *CESifo Forum*. Vol. 9. Ifo Institute for Economic Research at the University of Munich, 12–21.

[12.] Joh, S. W. 2003. Corporate governance and firm profitability: evidence from korea before the economic crisis. *Journal of Financial Economics 68*, 2, 287 – 322.

[13.] Kunz, T. 1993a. *Event Abstraction: Some Definitions and Theorems*. Citeseer.

[14.] Kunz, T. 1993b. Issues in event abstraction. In *PARLE '93 Parallel Architectures and Languages Europe*, A. Bode, M. Reeve, and G. Wolf, Eds. Lecture Notes in Computer Science, vol. 694. Springer Berlin Heidelberg, 668–671.

[15.] Li, Z., Wang, B., Li, M., and Ma, W.-Y. 2005. A probabilistic model for retrospective news event detection. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 106–113.

[16.] Luckham, D. 2003. The power of events: An introduction to complex event processing in distributed enterprise systems. *Ubiquity* .

[17.] Margara, A. and Cugola, G. 2011. Processing flows of information: from data stream to complex event processing. In *Proceedings of the 5th ACM international conference on Distributed event-based system*. ACM, 359–360.

[18.] Mitton, T. 2002. A cross-firm analysis of the impact of corporate governance on the east asian financial crisis. *Journal of financial economics 64*, 2, 215–241.

[19.] Nallapati, R., Feng, A., Peng, F., and Allan, J. 2004. Event threading within news topics. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. ACM, 446–453.

[20.] Obstfeld, M., Rogoff, K. S., Rogoff, K. S., and Rogoff, K. S. 2009. *Global imbalances and the financial crisis: products of common causes*. Centre for Economic Policy Research London.

[21.] Radev, D., Otterbacher, J., Winkel, A., and Blair-Goldensohn, S. 2005. Newsinessence: Summarizing online news topics. *Commun. ACM 48*, 10 (Oct.), 95–98. search, G. http://topforeignstocks.com/wp-content/uploads/2011/09/us-economic-crisis-timeline-300x163.jpg.

[22.] Soucy, P. and Mineau, G. W. 2005. Beyond tfidf weighting for text categorization in the vector space model. In *IJCAI*. Vol. 5. 1130–1135.

[23.] Steiner, S. 7 2013. Timeline of european debt crisis. *http://www.bankrate.com/finance/economics/timeline- european-debt-crisis-1.aspx. Accessed on*.

[24.] Vicente, K. 1990. The abstraction hierarchy as a basis for interface design: an empirical evaluation. In *Systems, Man and Cybernetics, 1990. Conference Proceedings., IEEE International Conference on*. 657 –659.

[25.] West, D. B. et al. 2001. *Introduction to graph theory*. Vol. 2. Prentice hall Upper Saddle River.

[26.] Yang, Y., Carbonell, J. G., Brown, R. D., Pierce, T., Archibald, B. T., and Liu, X. 1999. Learning approaches to topic detection and tracking.