Online Supplement: A Commonsense Knowledge-enabled Textual Analysis Approach for Financial Market Surveillance

Xin Li
Department of Information Systems, College of Business, City University of Hong Kong, Hong Kong Xin.Li.PhD@gmail.com,

Kun Chen
Department of Finance, South University of Science and Technology of China, Shenzhen, China, chen.k@sustc.cn,

Sherry X. Sun
sxsun2015@gmail.com,

Terrance Fung
Securities & Futures Commission of Hong Kong, Hong Kong, terrance.fung@gmail.com,

Huaiqing Wang
Department of Finance, South University of Science and Technology of China, Shenzhen, China, wang.hq@sustc.cn,

Daniel D. Zeng
Institute of Automation, Chinese Academy of Sciences, Beijing, China; Department of Management Information Systems, University of Arizona, Tucson, AZ, zeng@email.arizona.edu,

A. Related Research.

A.1. Financial Textual Analysis.

Analysis of textual market information is useful in finance, particularly in market trend prediction. In finance studies that investigate the impact of market information, Tetlock et al. (2008) found that the fraction of negative words in news can predict companies' earnings. Internet discussion boards can help predict market volatility (Antweiler and Frank 2004) and abnormal return (Tumarkin and Whitelaw 2001) after controlling for the effect of news. Xu and Zhang (2013) found that company announcements in Wikipedia weaken information asymmetry between managers and investors.

In data mining studies, Schumaker and Chen (2009) found that terms in news can provide extra predictive power for stock prices after factoring in autocorrelation of stock indices. Das and Chen (2007) also found that the volume and volatility of discussion board postings has predictive power for the overall market index. There have been quite a number

of studies exploring linguistic (Seo et al. 2003) and sentiment (Das and Chen 2007) features from news (Fung et al. 2003) and social media, such as microblogs (Oh and Sheng 2011, Ruiz et al. 2012), to predict stock price movements.

The efforts in financial text mining aid the development of automated trading systems (Mittermayer and Knolmayer 2006, Seo et al. 2003). However, financial text mining studies are focused on cumulative market response. These studies do not investigate individual transactions or untangle their connections with market information as in surveillance.

A.2. Prior Studies on Ontology-based Text Mining.

Commonsense knowledge has a long history in AI (Minsky 2000) and reasoning (Mueller 2009). In text mining, there are studies on extracting knowledge from natural language text and creating commonsense knowledge bases (Liu and Singh 2004, Richardson et al. 1998). In contrast to those studies, this study focuses on the application of commonsense knowledge and ontology in text mining.

One major application of ontology in text mining is feature disambiguation and generation. By employing knowledge bases, multiple terms in documents can be mapped to one concept. Less relevant terms, not shown in the ontology, can be removed from analysis. Both Cyc (Curtis et al. 2006) and WordNet (Seo et al. 2004) were used to differentiate ambiguous terms. In text summarization, Baralis et al. (2013) employed ontology for term disambiguation before using major sentences to summarize the document. In sentiment analysis, ontology has been used to restrict the analysis to certain aspects of product features (Penalver-Martinez et al. 2014, Zhou and Chaovalit 2008). In focused crawling, Zheng et al. (2008) applied neural networks on ontology concepts appearing in crawled Webpages as features to determine whether the links in the Webpages were worth further crawling.

Ontology is also employed to enrich linguistic features in text classification or clustering, where the antecedents and descendants of concepts appearing in documents are the generated new features (Gabrilovich and Markovitch 2007). Using WordNet (Fellbaum 1998), previous studies elaborated hypernymy, hyponymy, meronymy, and holonymy relations of words to enrich linguistic features (Bloehdorn et al. 2005, Chen et al. 2010, Hotho et al. 2003, Hung et al. 2004). Zheng et al. (2009) identified noun phrases in documents and their related words in WordNet for text clustering. Similar techniques are also applied on entities extracted from Wikipedia (Gabrilovich and Markovitch 2005, Gupta and Ratinov

2008). In sentiment analysis, Balahur et al. (2012) employed ontology to connect terms with their implied sentiments on texts without lexical clues.

From a text understanding perspective, in the two above streams of research, ontology helps decode the semantics of text. There are also studies exploiting ontology to connect related information for text mining, such as in query expansion.

Query expansion is a classic application of ontology in information retrieval (Jarvelin et al. 2001, Selvaretnam and Belkhatir 2012). The purpose is to help users provide more information about their queries and reduce the ambiguity. Bhogal et al. (2007) conducted a review on query expansion and classified it to relevance feedback, corpus dependent knowledge models, and corpus independent knowledge models, of which the latter two are ontology-based. Storey et al. (2008) took a corpus independent approach by combining commonsense knowledge bases with lexicons for query expansion. They match query terms with concepts in the ontology and used the terms' related concepts as expanded query terms. Since the extended terms may not match users' intention, user confirmation is usually needed (Storey et al. 2008). Jalali and Borujerdi (2011) combined relevance feedback and ontology-based methods by identifying terms frequently appearing in search engine's responses that were also concepts in the ontology to expand query terms. Han and Chen (2009) employed ontology to find similar users and cross-recommended terms among users for query expansion. Alipanah et al. (2012) employed the shortest path between the original term and expanded terms to weight the retrieved documents. In general, query expansion studies employ ontology and knowledge bases to fill in the missing information between user's original queries and the expanded queries.

B. Algorithm for Feature Presentation.

To reduce specialists' cognitive load, we do not directly show them the selected transactional features STF_i. Instead, we visualize the semantic relations between news and transactions as a graph, in which news articles are represented as nodes (with their titles and contents). This simplification is due to two concerns: 1) humans are better at understanding free text than textual features; 2) visualizing the network between news and transaction can help their reasoning. Due to feature selection, the presented news articles are a subset of RN_i. Table 1 shows the procedure to prune G and keep only the semantic relations and news related to STF_i for visualization.

Table 1 Algorithm for Pruning the Presented Graph.

Input: G, STF_i, F_{N_j} for each N_j in RN_i.
Output: PG_i – a graph containing news and semantic relations to be presented to specialists
set V_{PG} and E_{PG} as empty sets.
for each $p = (c \xrightarrow{\delta_1} c_1 ... \xrightarrow{\delta_n} c_n)$ in transactional feature $p \oplus f$ in STF_i:
for each N_j, if $c_n \in C_{N_j}$:
create an artificial node c' identified by N_j
replace c_n in p with c': $p = (c \xrightarrow{\delta_1} c_1 ... \xrightarrow{\delta_n} c')$
decompose p to a set of entities $V_p = \{c_i\}$ and their relations $E_p = \{(c_{i-1}, c_i)\}$
set $V_{PG} = V_{PG} \bigcup V_p$ and $E_{PG} = E_{PG} \bigcup E_p$
return $PG_i = (V_{PG}, E_{PG})$

C. Inter-coder Reliability.

To mimic real-world practice, we invited three domain experts to manually judge the risk of suspicious transactions based on their experience, public news, and other available transaction information. Each transaction is coded on a 4 point rating scale: "strongly disagree," "disagree," "agree," and "strongly agree" (on whether the transaction is high-risk). The scale does not have a "neither agree nor disagree" choice since this is not allowed in practice. A simple Web interface is used to input the ratings, and we provide a short individual training session on the use of the interface and the meaning of our scale. We do not provide a definition of high-risk transactions, since this is their expertise. The domain experts were asked to align their coding with what they do in their daily work, where high-risk transactions will be recommended for follow-up investigations. The domain experts were asked to conduct the coding independently; we did not observe them exchanging any information. (The domain experts received no monetary reward, so there was little incentive for them to cheat on this task.) On the raw coding results, the intra-class correlation coefficient (ICC) with two-way mixed effects on absolute agreement is 0.821, indicating an excellent agreement according to the commonly cited cutoffs provided by Cicchetti (1994) (Poor: <0.40; Fair: 0.40˜0.59; Good: 0.60˜0.74; Excellent: 0.75˜1.0). For the sake of our binary classification setup in this paper, we specify the "agree" and "strongly agree" ratings to be high-risk and the "strongly disagree" and "disagree" ratings to be normal. After this re-scaling, the ICC is 0.721, which is still a good agreement level. Noting that the re-scaled positive-negative coding may better be considered as nominal variables, we calculate the bias-adjusted kappa of (Siegel and Castellan 1988) . The kappa values of the pairs of the three domain experts are 0.547 [coders 1 and 2], 0.515 [coders 2 and 3], and 0.333 [coders 1 and 3], respectively. That means two pairs are at the level of

moderate agreement and one pair is at the level of fair agreement (Landis and Koch 1977) (Poor: <0; Slight: 0˜0.2; Fair: 0.21˜0.4; Moderate: 0.41˜0.6; Substantial: 0.61˜0.8; Perfect: 0.81˜1.0). Note that we asked the domain experts to code based on their practice, rather than using a coding schema defined by us (as in other empirical studies). Since all domain experts have extensive surveillance knowledge, we believe the inconsistent instances are due to the difficulty of the task, which, in practice, needs to be addressed through discussion.

D. Experimental Results.

D.1. Example Features.

Table 2 Feature Examples.

Methods	*Company*	*Entity Found in News*	*Transactional Features*		*News*
			Prefix	*News Features*	*Fea.Type*
NTR	AAC Technologies	Hardware Industry	Company→Industry	plunge, decline, loss ...	Linguistic
NTR+RE	AAC Technologies	Hardware Industry	Company→Industry	[revenueLost]...	Semantic
QE	AAC Technologies	Hardware Industry		plunge, decline, loss ...	Linguistic
QE+RE	AAC Technologies	Hardware Industry		[revenueLost]...	Semantic
NTR	Hutchison Harbour Ring Ltd.	Sir Li Ka-shing (through Hutchison Whampoa)	Company→Parent Company→KeyPerson	increase, stake ...	Linguistic
NM	BYD Electronic	BYD Electronic		profit, promotion, risk, invest, revenue...	Linguistic
NM+RE	BYD Electronic	BYD Electronic		[positive monetary income received by a business], [sale activity], [commercial activity]....	Semantic

D.2. Effect of Incomplete Knowledge Base.

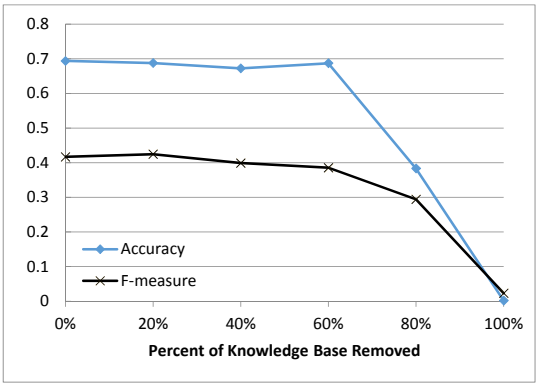


Figure 1 Performance of NTR+RE on the Optimistic Dataset on Incomplete Knowledge Bases.

A practical concern of using the commonsense knowledge-based approach is whether the knowledge base is comprehensive enough. Knowledge bases generally take years to

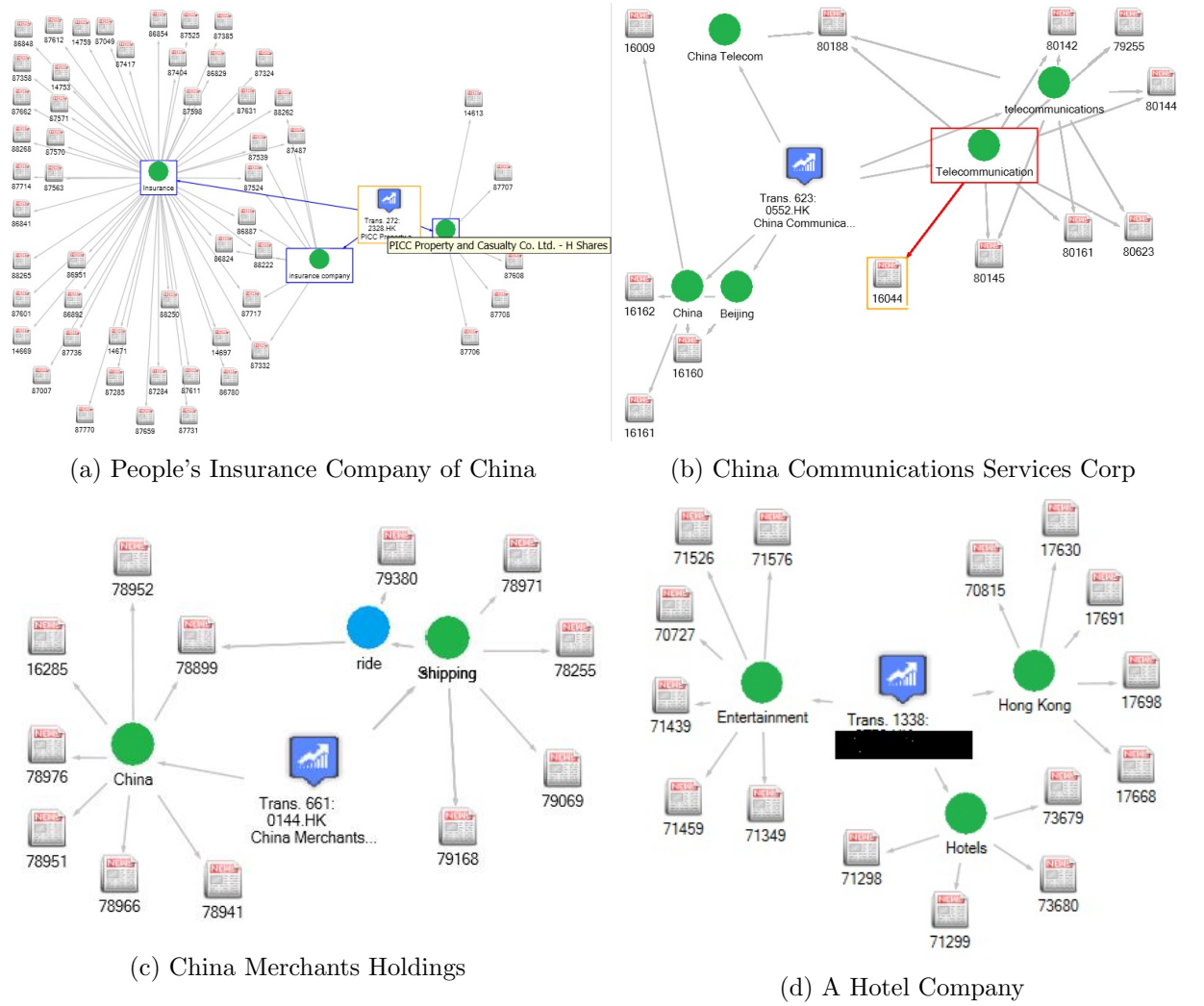
develop. If the knowledge base does not contain an entity, our proposed approach cannot generate its related features. To assess the impact of knowledge base quality on prediction performance, we conduct additional experiments by randomly removing entities from our knowledge base. Figure 1 reports the NTR+RE algorithm's results on the optimistic dataset when different percentages of entities are removed. We conduct 3 rounds of random entity removing for each threshold value. As we can see from the figure, when the knowledge base gets smaller, the prediction performance generally decreases. The decrease becomes faster when more entities are removed. Since the knowledge base we created is relatively comprehensive (for this research's purpose), the performance does not show a significant drop until 60% of the knowledge base is removed.

D.3. Case Studies.

Through the experiments, we find evidence that market information and market activity analyses provide complementary signals to support surveillance specialists' work. In some simple cases, the suspicious transaction can be directly explained by news mentioning the related company through name-entity search. Other scenarios, however, are more complicated.

Figure 2(a) shows a suspicious transaction on People's Insurance Company of China (PICC) that can be explained using indirectly related news. This transaction took place on September 16, 2008, when PICC's stock price experienced a significant drop. This price change is attributable to the then-current downtrend of the insurance industry. By jointly considering the industry sector news, industry stock performance, and company stock change, MarketWatch correctly predicted this transaction as a normal transaction.

Figure 2(b) provides another example highlighting the need for comprehensive commonsense knowledge bases. The transaction relates to China Communications Services Corporation Ltd. (CCSC), which is a subsidiary company of China Telecom, one of the four state-owned telecommunication providers in China. On October 8, 2008, CCSC's stock price dropped significantly. The MarketWatch system identified two major news articles indirectly related to this company on that day. The Ministry of Industry and Information Technology in China instituted a new regulation that required telecommunication companies to share their infrastructure for economic reasons (news #16044 in Figure 9(b)). The second news article (news #81088 in Figure 9(b)) projected that the new regulation would



(a) People's Insurance Company of China

(b) China Communications Services Corp

(c) China Merchants Holdings

(d) A Hotel Company

Figure 2 Cases Studies Using MarketWatch.

be useful for China Telecom, who was lagging behind in mobile services. Given this context, the price drop of CCSC stock appears highly suspicious, as predicted by the classifier. However, it should be noted that the commonsense knowledge of MarketWatch missed an important link: CCSC provides telecommunication infrastructure services to China Telecom. Due to collaboration with other companies, CCSC could lose its business when China Telecom downsizes its infrastructure plan. Thus, the stock price drop is a reasonable response to market information and does not require further investigation. Although the system needs improvement to avoid such mistakes, it successfully links related news and presents all decision cues together, minimizing surveillance specialists' efforts to identify the underlying reasons for the transaction.

Figure 2(c) provides an example of a high-risk transaction. The transaction took place on October 10, 2008, related to China Merchants Holdings (CMH), a large company in

logistics and financial service/investment. It was a black Friday on which the entire market experienced significant losses, including CMH's stock. The transaction occurred right before the market's close with a price about 30% higher than other transactions, at a volume of about 10% of that day's total volume. At the transaction time, no news article could explain such a change. Most news was discussing the economic crisis's negative impact on the industry. However, during the weekend, a rumor that the HK government would rescue the banking sector started to spread in the news. At the same time, the government confidently expressed that it would rescue the market through multiple channels. Due to this, as well as the Chinese government's follow-up announcement, CMH's major banks, CMB and Wing Lung Bank, experienced significant price increases in the following week, along with the entire sector. Accordingly, CMH's stock price pushed up to a much higher level. The buyer is suspected of receiving some insider news on the government's reaction before the market, and the transaction is worth further investigation.

Figure 2(d) is a transaction that experienced formal legal investigation[1]. The transaction involving a company in the hotel industry occurred at the end of October 2008. The stock had a low price and little activity. The transaction itself occupies about 30% of that day's volume and stock price increased by about 40%. A similar transaction occurred at the end of November at a similar volume. However, the market statuses were different when the two transactions occurred. In October, there were few news articles talking about the hotel industry. For the second one, there was a lot of mixed news about this industry and the trading was very active in this sector. (As a result, the second transaction pushed up the stock price on a smaller scale.) Through the MarketWatch system, the classifier is able to predict the risk of the first transaction. By drawing people's interest to the two highly similar transactions under quite different market contexts and with further supporting information, a human specialist would be reasonably concerned about the intention of the transactions in affecting stock price.

References

Alipanah, N., L. Khan, B. Thurisingham. 2012. Optimized ontology-driven query expansion using map-reduce framework to facilitate federated queries. *Computer Systems Science and Engineering* **27** 103–115.

Antweiler, W., M. Z. Frank. 2004. Is all that talk just noise? the information content of internet stock message boards. *Journal of Finance* **59** 1259–1294.

[1] Due to privacy concerns, we hide some information of this case on the figure and in the discussion.

Balahur, A., J. M. Hermida, A. Montoyo. 2012. Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems* **53** 742–753.

Baralis, E., L. Cagliero, S. Jabeen, A. Fiori, S. Shah. 2013. Multi-document summarization based on the yago ontology. *Expert Systems with Applications* **40** 6976–6984.

Bhogal, J., A. Macfarlane, P. Smith. 2007. A review of ontology based query expansion. *Information Processing & Management* **43** 866–886.

Bloehdorn, Stephan, Andreas Hotho, Steffen Staab. 2005. *An Ontology-based Framework for Text Mining*, *LDV Forum C GLDV Journal for computational linguistics and language technology*, vol. 20.

Chen, C. L., F. S. C. Tseng, T. Liang. 2010. An integration of word net and fuzzy association rule mining for multi-label document clustering. *Data & Knowledge Engineering* **69** 1208–1226.

Cicchetti, DV. 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* **6** 284C290.

Curtis, Jon, John Cabral, David Baxter. 2006. On the application of the cyc ontology to word sense disambiguation. *International Florida Artificial Intelligence Research Society Conference*. 652–657.

Das, S. R., M. Y. Chen. 2007. Yahoo! for amazon: Sentiment extraction from small talk on the web. *Management Science* **53** 1375–1388.

Fellbaum, V. 1998. *Introduction: WordNet: An Electronic Lexical DataBase*. MIT Press.

Fung, G.P.C., J.X. Yu, W. Lam. 2003. Stock prediction: Integrating text mining approach using real-time news. *IEEE International Conference on Computational Intelligence for Financial Engineering*. Hong Kong, 395–402.

Gabrilovich, E., S. Markovitch. 2007. Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization. *Journal of Machine Learning Research* **8** 2297–2345.

Gabrilovich, Evgeniy, Shaul Markovitch. 2005. Feature generation for text categorization using world knowledge. *International Joint Conferences on Artificial Intelligence*. 1048–1053.

Gupta, Rakesh, Lev Ratinov. 2008. Text categorization with knowledge transfer from heterogeneous data sources. *National Conference on Artificial Intelligence*. Chicago, Illinois.

Han, L. X., G. H. Chen. 2009. Hqe: A hybrid method for query expansion. *Expert Systems with Applications* **36** 7985–7991.

Hotho, Andreas, Steffen Staab, Gerd Stumme. 2003. Wordnet improves text document clustering. *Semantic Web Workshop of the 26th Annual International ACM SIGIR Conference*. Toronto, Canada.

Hung, C. L., S. Wermter, P. Smith. 2004. Hybrid neural document clustering using guided self-organization and wordnet. *IEEE Intelligent Systems* **19** 68–77.

Jalali, V., M. R. M. Borujerdi. 2011. Information retrieval with concept-based pseudo-relevance feedback in medline. *Knowledge and Information Systems* **29** 237–248.

Jarvelin, K., J. Kekalainen, T. Niemi. 2001. Expansiontool: Concept-based query expansion and construction. *Information Retrieval* **4** 231–255.

Landis, JR, GG Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* **33** 159–174.

Liu, H., P. Singh. 2004. Concept net - a practical commonsense reasoning tool-kit. *BT Technology Journal* **22** 211–226.

Minsky, M. 2000. Commonsense-based interfaces. *Communications of the ACM* **43** 67–73.

Mittermayer, M.A., G.F. Knolmayer. 2006. Newscats: A news categorization and trading system. *International Conference in Data Mining*. Hong Kong.

Mueller, E. T. 2009. Automating commonsense reasoning using the event calculus. *Communications of the ACM* **52** 113–117.

Oh, Chong, Olivia Sheng. 2011. Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. *International Conference on Information Systems*. Shanghai, 17.

Penalver-Martinez, I., F. Garcia-Sanchez, R. Valencia-Garcia, M. A. Rodriguez-Garcia, V. Moreno, A. Fraga, J. L. Sanchez-Cervantes. 2014. Feature-based opinion mining through ontologies. *Expert Systems with Applications* **41** 5995–6008.

Richardson, S. D., W. B. Dolan, L. Vanderwende. 1998. Mindnet: acquiring and structuring semantic information from text. *the 17th International Conference on Computational linguistics*, vol. 2. 1098–1102.

Ruiz, Eduardo J., Vagelis Hristidis, Carlos Castillo, Aristides Gionis, Alejandro Jaimes. 2012. Correlating financial time series with micro-blogging activity. *ACM International Conference on Web Search and Data Mining*.

Schumaker, R. P., H. C. Chen. 2009. Textual analysis of stock market prediction using breaking financial news: The azfintext system. *ACM Transactions on Information Systems* **27**.

Selvaretnam, B., M. Belkhatir. 2012. Natural language technology and query expansion: issues, state-of-the-art and perspectives. *Journal of Intelligent Information Systems* **38** 709–740.

Seo, H. C., H. J. Chung, H. C. Rim, S. H. Myaeng, S. H. Kim. 2004. Unsupervised word sense disambiguation using wordnet relatives. *Computer Speech and Language* **18** 253–273.

Seo, Y.W., J.A. Giampapa, K.P. Sycara. 2003. Financial news analais for intelligent portfolio management. *American Association for Artificial Intelligence*.

Siegel, S, NJ. Castellan. 1988. *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York.

Storey, V. C., A. Burton-Jones, V. Sugurnaran, S. Purao. 2008. Conquer: A methodology for context-aware query processing on the world wide web. *Information Systems Research* **19** 3–25.

Tetlock, P. C., M. Saar-Tsechansky, S. Macskassy. 2008. More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance* **63** 1437–1467.

Tumarkin, R., R. F. Whitelaw. 2001. News or noise? internet postings and stock prices. *Financial Analysts Journal* **57** 41–51.

Xu, Sean Xin, Xiaoquan (Michael) Zhang. 2013. Impact of wikipedia on market information environment: Evidence on management disclosure and investor reaction. *MIS Quarterly* **37** 1043–1068.

Zheng, H. T., B. Y. Kang, H. G. Kim. 2008. Learnable focused crawling based on ontology. *Information Retrieval Technology* **4993** 264–275.

Zheng, H. T., B. Y. Kang, H. G. Kim. 2009. Exploiting noun phrases and semantic relationships for text document clustering. *Information Sciences* **179** 2249–2262.

Zhou, L., P. Chaovalit. 2008. Ontology-supported polarity mining. *Journal of the American Society for Information Science and Technology* **59** 98–110.