A Computationally Efficient Measure for Word Semantic Relatedness Using Time Series

Arash Joorabchi[1], Alaa Alahmadi[1], Michael English[2], Abdulhussain E. Mahdi[1]
[1]Department of Electronic and Computer Engineering
[2]Department of Computer Science and Information Systems
University of Limerick
Limerick, Ireland
{arash.joorabchi, alaa.alahmadi, michael.english, hussain.mahdi}@ul.ie

***Abstract*—Measurement of words semantic relatedness plays an important role in a wide range of natural language processing and information retrieval applications, such as full-text search, summarization, classification and clustering. In this paper, we propose an easy to implement and low-cost method for estimating words semantic relatedness. The proposed method is based on the utilization of words temporal footprints as found in publicly available corpora such as Google Books Ngrams (GBN), and knowledge bases such as Wikipedia. The extracted footprints are represented as time series, their similarities is measured using the Minkowski distance, and averaged using a correlation-based weighting scheme to quantify the words semantic relatedness. The overall performance of the method and the quality of the two sources used for extracting words temporal footprints (i.e., GBN and Wikipedia) are evaluated using the MTurk-287 dataset and the standard measures of Pearson's *r* and Spearman's *ρ*.**

Keywords—Word semantic relatedness; time series; temporal features

I. INTRODUCTION

The task of quantifying Words Semantic Relatedness (WSR) is a fundamental building block of various Natural Language Processing (NLP) and Information Retrieval (IR) systems. The goal of this task is to estimate the semantic distance between a given pair of words as close to that estimated by humans. Examples of NLP and IR systems which rely on accurate computation of WSR include: Word Sense Disambiguation (WSD) [1], document clustering [2], search query optimization [3], text summarization [4], and evaluation of machine translation [5]. Methods developed to measure WSR can be divided into two main categories of corpus-based and knowledge-based [6, 7]. Corpus-based methods utilize large corpora to estimate the relatedness of words based on various statistical criteria such as the probability of their co-occurrence. Well-known examples of these methods include: Pointwise Mutual Information (PMI) [8] and Latent Semantic Analysis (LSA) [9].

Knowledge-based methods take advantage of the semantic information encoded (by humans) in lexical databases such as WordNet [10] and knowledge bases such as Wikipedia. For example, Leacock and Chodorow [11] proposed a WSR method which estimates the relatedness of a pair of words/concepts in WordNet based on the distance (i.e., normalized length of the shortest path) between them as found in the WordNet is-a hierarchy graph. Another example of WordNet based methods is the work of Banerjee and Pedersen [12] which measures the semantic relatedness of two WordNet concepts based on the level of overlap (shared words) between their definitions (glosses). The performance of knowledge-based methods using WordNet is limited by the relatively small size of this knowledge base (currently 117,000 concepts). This limitation of WordNet has led to the use of Wikipedia as an alternative knowledge base. The English Wikipedia currently contains over 4 million articles/concepts covering subjects in all aspects of human knowledge and growing. This makes Wikipedia one of the most comprehensive knowledge bases currently available. The wide coverage of Wikipedia along with its up-to-datedness (due to its crowd-sourced nature), rich semantics, and multilingual nature make it an effective knowledge base for building knowledge-based WSR methods.

Two well-known examples of knowledge-based methods using Wikipedia are WikiRelate [13] and Wikipedia Link-based Measure (WLM) [14]. Wikipedia articles are classified according to the Wikipedia's own community-built classification scheme. This scheme has a loose semi-hierarchical directed-graph structure which allows articles to belong to multiple categories, and categories to have multiple parent categories (currently going up to 16 levels of depth). Utilizing this feature of Wikipedia, the WikiRelate method estimates the relatedness of two Wikipedia articles/concepts based on the normalized length of the shortest path between them as found in the Wikipedia's classification graph. Wikipedia articles are inter-connected via an intricate network of hyperlinks which can be mined for discovering associative relations between the represented concepts. The WLM method utilizes this network to quantify the relatedness of two concepts. In this method the relatedness between two Wikipedia articles/concepts is measured based on the number of Wikipedia concepts which discuss/mention and have hyperlinks to both the two concepts being compared.

Explicit Semantic Analysis (ESA) [15] is another example of knowledge-based WSR methods using Wikipedia which outperforms both WikiRelate and WLM. Unlike the earlier

methods which were based on utilizing the Wikipedia's classification graph and inter-article networks, ESA uses the textual content of Wikipedia articles directly in a vector space model. In this method each word is mapped to a vector of Wikipedia articles (concepts) in which it appears and the entries in the vector contain the weights (TFIDF) of the word in those articles. The relatedness of a pair of words is then quantified by measuring the cosine similarity of their vectors. Temporal Semantic Analysis (TSA) [16] is a temporally enhanced version of ESA which has achieved the state-of-the-art performance in WSR. The TSA is based on the premise that the temporal information of words may be used as a complementary signal for measuring WSR. For example, similar occurrence rates of the words "*war*" and "*peace*" over time could signal their relatedness. The TSA algorithm mines this temporal information from a historical archive (New York Times articles published since 1870) and uses them to complement the vector space model of ESA, such that each entry in the vector contains the time series of the corresponding concept rather than its TFIDF weight. The TSA estimates the relatedness of a pair of words by measuring the distance between their vectors of concept time series.

In this work we propose a simple and easy-to-implement method for measuring WSR which relies solely on the words' temporal characteristics as extracted from two independent sources, i.e., Google Books Ngrams and Wikipedia. We have evaluated the performance of the proposed method when using these sources individually and combined.

The rest of the paper is organized as follows: Section 2 describes the proposed temporal-based WSR method and its implementation details. Section 3 describes the evaluation criteria and the test datasets; and presents the results. This is followed by Section 4 which provides a conclusion and discusses future work.

II. TEMPORAL-BASED WSR

The goal of the proposed Temporal-based WSR (TWSR) method is to put forward a simple approach for measuring words semantic relatedness solely based on their temporal footprints. This approach works with words directly without mapping them to their corresponding concepts in a knowledge base, and therefore avoids the complexities and overload arising from such mapping process, e.g., the need for word sense disambiguation. This differentiates the TWSR from similar approaches, such as TSA, which use the temporal information as a complementary signal to enrich the vector of concepts.

A. Words Temporal Data Sources, Retrieval, and Normalization

In this work we have used two independent sources to acquire words temporal information, namely Google Books Ngrams (GBN) [17] and Wikipedia Page Views (WPV) statistics.

The GBN corpora are built based on the content of over 8 million books published from 1500 to 2008. The English GBN corpus contains about half a trillion words and captures their annual occurrence frequency in 4.5 million digitalized books over a span of 508 years [18]. Only the words which appear in at least 40 books are included in the corpus and the frequency counts are normalized by the number of books published in each year. Using this corpus we can build a 508-point time series for virtually any word, reflecting its rate of usage in books published in half a millennium. The GBN corpus is accessible via the Google's Ngram Viewer[1]; the corpus may be downloaded in bulk or, alternatively, HTML queries for individual words could be submitted, returning HTML pages containing the words time series in JSON format.

We use Wikipedia Page Views (WPV) as a second source for acquiring words temporal data. Since December 2007, Wikipedia has been gathering and publishing its page view count statistics. This includes counting the hourly views of the title of article pages and redirect pages. The English Wikipedia currently includes about 4.8 million unique article pages and 7 million redirect pages. Therefore, the WPV corpus could be used as a comprehensive source of words temporal information. The corpus can be either downloaded in bulk[2] or, alternatively, HTML queries for individual words could be submitted to an interface to the corpus which would return the words time series in JSON format[3]. Also, recently (end of 2015), Wikipedia released an API for accessing the WPV corpus[4]. In this work, we have used the article and redirect pages daily counts from December 2007 to December 2015. The resulting time series cover a span of 8 years with 2,923 time points.

As the final step of the data acquisition process, we standardize (normalize) the words GBN and WPV time-frequency time series by converting their raw frequency values to their corresponding z-scores such that:

$$z_i = \frac{x_i - \mu}{\sigma} \quad (1)$$

where, μ is the mean of the raw frequencies in the time series $X=\{x_1, x_2, \ldots, x_n\}$and σ is their standard deviation. This eliminates the unwanted discrepancies in the words raw frequency values due to their level of generality/specificity (usage scale). For example, the time series for the words "*hard drive*" and "*computer*" are similar in shape (i.e., correlate) but are different in scale, as the latter word is more generic than the former and used more often. This type of scale discrepancies could have a negative effect depending on the measure used to quantify the distance between the words time series. Figures 1&2 show the standardized GBN and WPV time series for a sample pair of words.

[1] https://books.google.com/ngrams
[2] http://dumps.wikimedia.org/other/pagecounts-raw/
[3] http://stats.grok.se/
[4] https://wikitech.wikimedia.org/wiki/Analytics/PageviewAPI

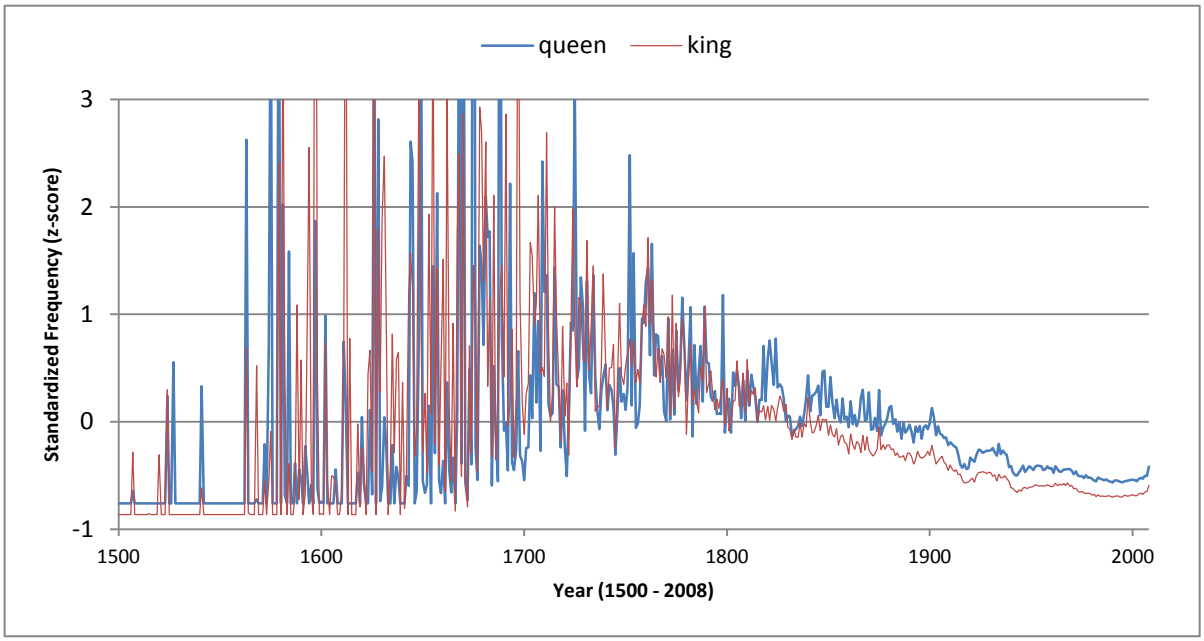

Fig. 1. Sample GBN time series (Pearson's r = 0.6)

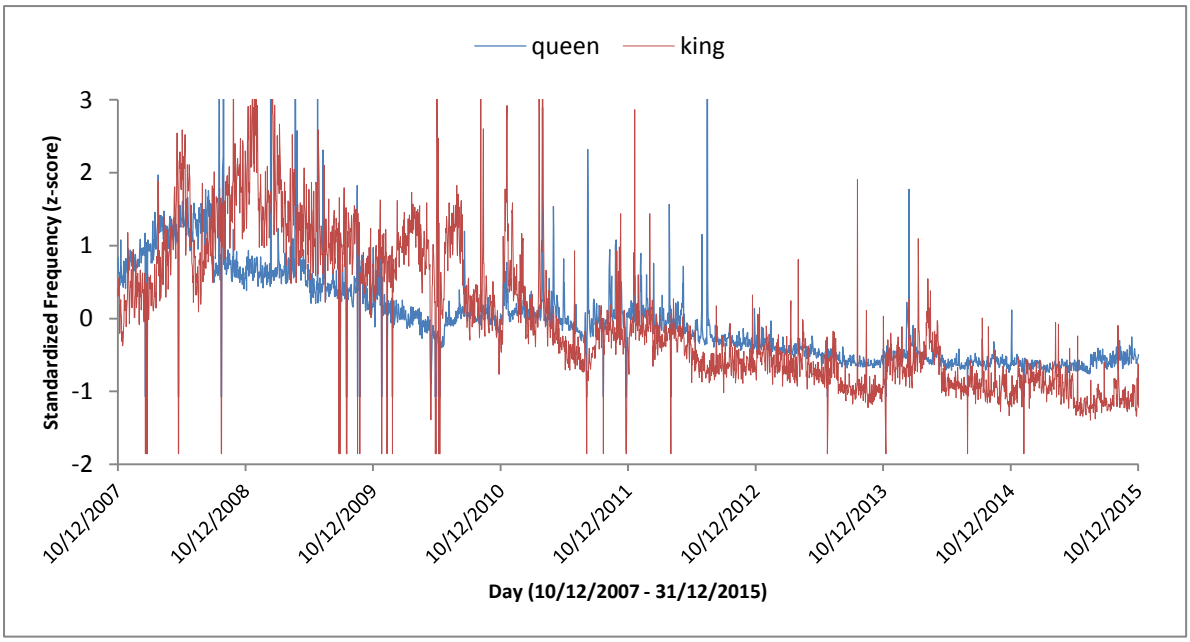

Fig. 2. Sample WPV time series (Pearson's r = 0.5)

B. Measuring Words Time Series Distances

Given a pair of words, we measure the distance/similarity between their corresponding GBN and WPV time series to quantify their relatedness. We experimented with various time series similarity measures and distance metrics including: cosine similarity, Euclidean distance, Manhattan distance, Minkowski distance, Pearson's and Spearman's correlation coefficients, and dynamic time warping [19]. The preliminary experiments showed the Minkowski distance (a.k.a. Lp-norm) to be the most suitable metric for this task. The Minkowski distance of order p between two time series X= $\{x_1, x_2 \ldots x_n\}$ and Y= $\{y_1, y_2 \ldots y_n\}$ is defined as:

$$d(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p} \quad (2)$$

The Minkowski distance is the generalization of the well-known Euclidean distance (p=2) and Manhattan distance (p=1). Empirically, we found the p value of 1.6 to yield the best results for this task.

Applying the Minkowski distance to the time series of each pair of words results in two independent distance values, one based on GBN and the other based on WPV. At this point, we can take either of these distance values as the relatedness value of the given words, or combine the two values by

averaging. We have experimented with both a simple mean average and a weighted average. In the weighted average approach, we first evaluate the Spearman's correlation of GBN and WPV-based relatedness values with that assigned by humans in a test dataset. We then use these correlation coefficient values as weights for the GBN and WPV distance values when averaging.

III. Evaluation

We have used the MTurk-287 dataset compiled by Radinsky et al. [16] to evaluate the performance of the proposed TWSR method and compare it with that of the state-of-the-art methods. The dataset contains a total of 287 pairs of words. The pairs are chosen with the goal of creating a dataset with a balanced number of related and unrelated words. The relatedness of each pair is evaluated and scored by 10 different individuals. The Human assigned scores for each pair are averaged to produce a single relatedness value to be used as the ground truth.

Following the literature, we have used the Pearson's correlation coefficient r and Spearman's correlation coefficient ρ to measure the level of correlation between the relatedness scores assigned by the proposed method and humans. Table 1 presents the evaluation results; and Table 2 shows the first 20 word pairs from the dataset along with their human- and machine-assigned relatedness scores[5].

The evaluation results show that the WPV time series provide a stronger signal than the GBN time series (ρ=0.32 vs. ρ=0.29). This is an interesting finding considering the fact that the GBN series cover a much longer time span than the WPV series (508 years vs. 8 years); whereas, the WPV series are more fine-grained and have much more time points than the GBN series (2,923 vs. 508). The TWSR method achieves the best results when the GBN and WPV signals are combined using a weighted average. The accuracy performance of the TWSR falls short from the current state-of-the-art method: TSA (ρ=0.40 vs. ρ=0.63). However, we believe its simplicity and low computation cost make it a viable alternative to more complex WSR methods with a higher accuracy. The relatively lower accuracy of the TWSR can be contributed to the fact that it does not address the issue of words sense ambiguity. For example, given the sample pair of words "plane" and "aircraft", their relatedness score could dramatically change depending on the intended sense of the word "plane", e.g., plane (Fixed-wing aircraft) vs. plane (geometry). Since the Wikipedia articles/concepts are disambiguated (i.e., there are separate articles for different senses of a word), the TWSR method could be enhanced to consider different senses of words when measuring their relatedness. In its current form, the TWSR compares the WPV time series of the most commonly used senses of the given pair of words to measure their relatedness. However, its enhanced version would compare the WPV time series of all senses of the pair to find the most related ones and infer them as the intended senses of the words.

[5] The full dataset along with the WPV and GBNS time series of its 500 unique words are available at: http://www.skynet.ie/~arash/zip/TWSR_v1.zip

TABLE I. Evaluation resutls

Method		Correlation with Humans	
		Pearson's r	***Spearman's ρ***
TWSR	GBN time series	0.27	0.29
	WPV time series	0.33	0.32
	GBN + WPV (mean)	0.38	0.39
	GBN + WPV (weighted average)	0.39	0.40
ESA [15]		n/a	0.59
TSA [16]		n/a	0.63

TABLE II. Sample word pairs from the MTurk-287 dataset and their relatedness scores.

Word Pair	Relatedness Scores (0-10)			
	Humans	***WPV***	***GBN***	***WPV + GBN (weighted Average)***
episcopal , russia	4.07	4.02	2.19	2.86
water , shortage	3.97	2.00	2.93	2.27
horse , wedding	2.61	4.45	2.90	3.56
plays , losses	5.43	4.99	4.43	4.86
classics , advertiser	2.56	4.78	3.34	4.03
latin , credit	2.00	3.69	3.48	3.54
ship , ballots	2.75	2.79	2.22	2.22
mistake , error	8.91	7.08	5.06	6.41
disease , plague	8.20	6.66	1.79	4.01
sake , shade	3.41	6.76	4.91	6.13
saints , observatory	1.62	3.46	1.42	2.05
treaty , wheat	1.24	2.63	3.10	2.72
texas , death	0.40	7.10	0.27	3.23
republicans , challenge	2.75	2.88	2.73	2.60
body , peaceful	1.99	4.03	2.07	2.78
admiralty , intensity	3.76	4.59	2.68	3.49
body , improving	2.17	2.79	3.69	3.19
heroin , marijuana	5.96	5.58	9.58	8.61
scottish , commuters	3.88	2.79	2.58	2.46
apollo , myth	3.62	2.40	2.78	2.38

IV. Conclusion

In this paper we investigated the application of words temporal data for measuring their semantic relatedness, and proposed a simple method, TWSR, which uses words temporal data acquired from two independent sources, namely Wikipedia and Google Books Ngrams, to quantify words semantic relatedness. We evaluated the accuracy performance of the proposed method using the MTurk-287 dataset and standard measures of Pearson's r and Spearman's ρ correlation coefficients. The accuracy of the TWSR falls short from that of the state-of-the-art method; however, it provides an easy-to-implement and low computational cost alternative to more

complex methods such as ESA and TSA. Also, the results of reported experiments with the WPV time series show that they may be directly used as an independent temporal feature to enhance the knowledge-based WSR methods using Wikipedia, such as WikiRelate [13] and Wikipedia Link-based Measure (WLM) [14].

As discussed in Section 3, the performance of the TWSR may be further improved by addressing the issue of words sense ambiguity. Therefore, as future work, we plan to develop and evaluate a new version of the TWSR which would consider all possible senses of a given pair of words (instead of the most common ones) and compare their corresponding time series to find the ones with the lowest distance as the right senses for the given words. Also, an interesting avenue for future research is to explore the possibility of using temporal time series for developing a dynamic temporal-based word relatedness measure, where a given pair of words would be assigned different relatedness scores for different time periods.

REFERENCES

[1] Patwardhan S., Banerjee S. and Pedersen T. Using measures of semantic relatedness for word sense disambiguation. In: Proceedings of the 4th international conference on Computational linguistics and intelligent text processing; 2003; Mexico City, Mexico: Springer-Verlag; 2003. p. 241-257.

[2] Yazdani M. and Popescu-Belis A. Using a Wikipedia-based semantic relatedness measure for document clustering. In: Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing; 2011; Portland, Oregon: Association for Computational Linguistics; 2011. p. 29-36.

[3] Freitas A., Oliveira J. G., O'Riain S., da Silva J. C. P. and Curry E., Querying linked data graphs using semantic relatedness: A vocabulary independent approach, Data & Knowledge Engineering 2013; 88: 126-141.

[4] Shasha X. and Yang L. Using corpus and knowledge-based similarity measure in Maximum Marginal Relevance for meeting summarization. In: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on; 2008 March 31 2008-April 4 2008; 2008. p. 4985-4988.

[5] Padó S., Cer D., Galley M., Jurafsky D. and Manning C. D., Measuring machine translation quality as semantic equivalence: A metric based on entailment features, Machine Translation 2009; 23, 2: 181-193.

[6] Gomaa W. H. and Fahmy A. A., A survey of text similarity approaches, International Journal of Computer Applications 2013; 68, 13.

[7] Mihalcea R., Corley C. and Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity. In: Proceedings of the 21st national conference on Artificial intelligence - Volume 1; 2006; Boston, Massachusetts: AAAI Press; 2006. p. 775-780.

[8] Turney P. D. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In: L. Raedt and P. Flach, (eds.). Machine Learning: ECML 2001: 12th European Conference on Machine Learning Freiburg, Germany, September 5–7, 2001 Proceedings. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, p. 491-502.

[9] Landauer T. K., Foltz P. W. and Laham D., An introduction to latent semantic analysis, Discourse Processes 1998; 25, 2-3: 259-284.

[10] Miller G. A., WordNet: a lexical database for English, Commun. ACM 1995; 38, 11: 39-41.

[11] Leacock C. and Chodorow M. Combining local context and WordNet similarity for word sense identification. WordNet: An Electronic Lexical Database. In C. Fellbaum (Ed.), MIT Press, 1998, p. 265-283.

[12] Banerjee S. and Pedersen T. Extended gloss overlaps as a measure of semantic relatedness. In: Ijcai; 2003; 2003. p. 805-810.

[13] Strube M. and Ponzetto S. P. WikiRelate! computing semantic relatedness using wikipedia. In: proceedings of the 21st national conference on Artificial intelligence - Volume 2; 2006; Boston, Massachusetts: AAAI Press; 2006. p. 1419-1424.

[14] Milne D. and Witten I. H. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In: first AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08); 2008; Chicago, I.L; 2008.

[15] Gabrilovich E. and Markovitch S. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In: Proceedings of the 20th international joint conference on Artifical intelligence; 2007; Hyderabad, India: Morgan Kaufmann Publishers Inc.; 2007. p. 1606-1611.

[16] Radinsky K., Agichtein E., Gabrilovich E. and Markovitch S. A word at a time: computing word relatedness using temporal semantic analysis. In: Proceedings of the 20th international conference on World wide web; 2011; Hyderabad, India: ACM; 2011. p. 337-346.

[17] Michel J.-B., Shen Y. K., Aiden A. P., Veres A., Gray M. K., Pickett J. P., Hoiberg D., Clancy D., Norvig P., Orwant J., Pinker S., Nowak M. A. and Aiden E. L., Quantitative Analysis of Culture Using Millions of Digitized Books, Science 2011; 331, 6014: 176-182.

[18] Lin Y., Michel J.-B., Aiden E. L., Orwant J., Brockman W. and Petrov S. Syntactic annotations for the Google Books Ngram Corpus. In: Proceedings of the ACL 2012 System Demonstrations; 2012; Jeju Island, Korea: Association for Computational Linguistics; 2012. p. 169-174.

[19] Cassisi C., Pulvirenti A., Cannata A., Aliotta M. and Montalto P., Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining (INTECH Open Access Publisher, 2012).