Detection of Topic and its Extrinsic Evaluation Through Multi-Document Summarization

Yoshimi Suzuki
Interdisciplinary Graduate School of
Medicine and Engineering
University of Yamanashi
Kofu, 400-8511, JAPAN
ysuzuki@yamanashi.ac.jp

Fumiyo Fukumoto
Interdisciplinary Graduate School of
Medicine and Engineering
University of Yamanashi
Kofu, 400-8511, JAPAN
fukumoto@yamanashi.ac.jp

Abstract

This paper presents a method for detecting words related to a topic (we call them topic words) over time in the stream of documents. Topic words are widely distributed in the stream of documents, and sometimes they frequently appear in the documents, and sometimes not. We propose a method to reinforce topic words with low frequencies by collecting documents from the corpus, and applied Latent Dirichlet Allocation (Blei et al., 2003) to these documents. For the results of LDA, we identified topic words by using Moving Average Convergence Divergence. In order to evaluate the method, we applied the results of topic detection to extractive multi-document summarization. The results showed that the method was effective for sentence selection in summarization.

1 Introduction

As the volume of online documents has drastically increased, the analysis of topic bursts, topic drift or detection of topic is a practical problem attracting more and more attention (Allan et al., 1998; Swan and Allan, 2000; Allan, 2003; Klinkenberg, 2004; Lazarescu et al., 2004; Folino et al., 2007). The earliest known approach is the work of Klinkenberg and Joachims (Klinkenberg and Joachims, 2000). They have attempted to handle concept changes by focusing a window with documents sufficiently close to the target concept. Mane *et. al.* proposed a method to generate maps that support the identification of major research topics and trends (Mane and Borner, 2004). The method used Kleinberg's burst detection algorithm, co-occurrences of words, and graph layout technique. Scholz *et. al.* have attempted to use different ensembles obtained by training several data streams to detect concept drift (Scholz, 2007). However the ensemble method itself remains a problem that how to manage several classifiers effectively. He and Parket attempted to find bursts, periods of elevated occurrence of events as a dynamic phenomenon instead of focusing on arrival rates (He and Parker, 2010). However, the fact that topics are widely distributed in the stream of documents, and sometimes they frequently appear in the documents, and sometimes not often hamper such attempts.

This paper proposes a method for detecting topic over time in series of documents. We reinforced words related to a topic with low frequencies by collecting documents from the corpus, and applied Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to these documents in order to extract topic candidates. For the results of LDA, we applied Moving Average Convergence Divergence (MACD) to find topic words while He *et. al.*, applied it to find bursts. The MACD is a technique to analyze stock market trends (Murphy, 1999). It shows the relationship between two moving averages of prices modeling bursts as intervals of topic dynamics, *i.e.*, positive acceleration. Fukumoto *et. al* also applied MACD to find topics. However, they applied it only to the words with high frequencies in the documents (Fukumoto et al., 2013). In contrast, we applied it to the topic candidates obtained by LDA.

We examined our method by extrinsic evaluation, *i.e.*, we applied the results of topic detection to extractive multi-document summarization. We assume that a salient sentence includes words related to the target topic, and an event of each documents. Here, an event is something that occurs at a specific place and time associated with some specific actions(Allan et al., 1998). We identified event words by using the traditional tf$*$idf method applied to the results of named entities. Each sentence in documents is represented using a vector of frequency weighted words that can be event

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 241–246,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

or topic words. We used Markov Random Walk (MRW) to compute the rank scores for the sentences (Page et al., 1998). Finally, we selected a certain number of sentences according to the rank score into a summary.

2 Topic Detection

2.1 Extraction of Topic Candidates

LDA presented by (Blei et al., 2003) models each document as a mixture of topics (we call it lda_topic to discriminate our *topic* candidates), and generates a discrete probability distribution over words for each lda_topic. The generative process for LDA can be described as follows:

1. For each topic k = 1, $\cdots$, K, generate ϕ_k, multinomial distribution of words specific to the topic k from a Dirichlet distribution with parameter β;

2. For each document d = 1, $\cdots$, D, generate θ_d, multinomial distribution of topics specific to the document d from a Dirichlet distribution with parameter α;

3. For each word n = 1, $\cdots$, N_d in document d;
   (a) Generate a topic z_{dn} of the n^{th} word in the document d from the multinomial distribution θ_d
   (b) Generate a word w_{dn}, the word associated with the n^{th} word in document d from multinomial ϕ_{zdn}

Like much previous work on LDA, we used Gibbs sampling to estimate ϕ and θ. The sampling probability for topic z_i in document d is given by:

$$P(z_i \mid z_{\setminus i}, W) = \frac{(n^v_{\setminus i,j}+\beta)(n^d_{\setminus i,j}+\alpha)}{(n^{\cdot}_{\setminus i,j}+W\beta)(n^d_{\setminus i,\cdot}+T\alpha)}. \quad (1)$$

$z_{\setminus i}$ refers to a topic set Z, not including the current assignment z_i. $n^v_{\setminus i,j}$ is the count of word v in topic j that does not include the current assignment z_i, and $n^{\cdot}_{\setminus i,j}$ indicates a summation over that dimension. W refers to a set of documents, and T denotes the total number of unique topics. After a sufficient number of sampling iterations, the approximated posterior can be used to estimate ϕ and θ by examining the counts of word assignments to topics and topic occurrences in documents. The approximated probability of topic k in the document d, $\hat{\theta}^k_d$, and the assignments word w to topic k, $\hat{\phi}^w_k$ are given by:

$$\hat{\theta}^k_d = \frac{N_{dk}+\alpha}{N_d+\alpha K}. \quad (2)$$

$$\hat{\phi}^w_k = \frac{N_{kw}+\beta}{N_k+\beta V}. \quad (3)$$

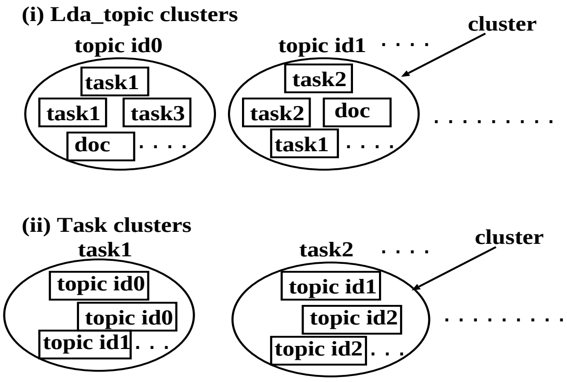

Figure 1: Lda_topic cluster and task cluster

We used documents prepared by summarization tasks, NTCIR and DUC data as each task consists of series of documents with the same topic. We applied LDA to the set consisting of all documents in the summarization tasks and documents from the corpus. We need to estimate the appropriate number of lda_topic.

Let k' be the number of lda_topics and d' be the number of topmost d' documents assigned to each lda_topic. We note that the result obtained by LDA can be regarded as the two types of clustering result shown in Figure 1: (i) each cluster corresponds to each lda_topic (topic id0, topic id1 $\cdots$ in Figure 1), and each element of the clusters is the document in the summarization tasks (task1, task2, $\cdots$ in Figure 1) or from the corpus (doc in Figure 1), and (ii) each cluster corresponds to the summarization task and each element of the clusters is the document in the summarization tasks or the document from the corpus assigned topic id. For example, DUC2005 consists of 50 tasks. Therefore the number of different clusters is 50. We call the former lda_topic cluster and the latter task cluster. We estimated k' and d' by using Entropy measure given by:

$$E = -\frac{1}{\log l}\sum_j \frac{N_j}{N}\sum_i P(A_i,C_j)\log P(A_i,C_j) \quad (4)$$

l refers to the number of clusters. $P(A_i, C_j)$ is a probability that the elements of the cluster C_j assigned to the correct class A_i. N denotes the total number of elements and N_j shows the total number of elements assigned to the cluster C_j. The value of E ranges from 0 to 1, and the smaller value of E indicates better result. Let E_{topic} and E_{task} are entropy value of lda_topic cluster and task cluster, respectively. We chose the parameters k' and d' whose value of the summation of E_{topic} and E_{task} is smallest. For each lda_topic, we extracted words whose probabilities are larger than zero, and regarded these as topic candidates.

2.2 Topic Detection by MACD

The proposed method does not simply use MACD to find bursts, but instead determines topic words in series of documents. Unlike Dynamic Topic Models (Blei and Lafferty, 2006), it does not assume Gaussian distribution so that it is a natural way to analyze bursts which depend on the data. We applied it to extract topic words in series of documents. MACD histogram defined by Eq. (6) shows a difference between the MACD and its moving average. MACD of a variable x_t is defined by the difference of n_1-day and n_2-day moving averages, MACD(n_1,n_2) = EMA(n_1) - EMA(n_2). Here, EMA(n_i) refers to n_i-day Exponential Moving Average (EMA). For a variable $x = x(t)$ which has a corresponding discrete time series x = $\{x_t \mid t$ = 0,1,$\cdots\}$, the n-day EMA is defined by Eq. (5).

$$\begin{aligned}\text{EMA}(n)[x]_t &= \alpha x_t + (1-\alpha)\text{EMA}(n-1)[x]_{t-1} \\ &= \sum_{k=0}^{n} \alpha(1-\alpha)^k x_{t-k}. \qquad (5)\end{aligned}$$

α refers to a smoothing factor and it is often taken to be $\frac{2}{(n+1)}$. MACD histogram shows a difference between the MACD and its moving average[1].

$$\begin{aligned}\text{hist}(n_1,n_2,n_3) &= \text{MACD}(n_1,n_2) - \\ &\quad \text{EMA}(n_3)[\text{MACD}(n_1,n_2)]. \qquad (6)\end{aligned}$$

The procedure for topic detection with MACD is illustrated in Figure 2. Let A be a series of documents and w be one of the topic candidates obtained by LDA. Each document in A is sorted in chronological order. We set A to the documents from the summarization task. Whether or not a word w is a topic word is judged as follows:

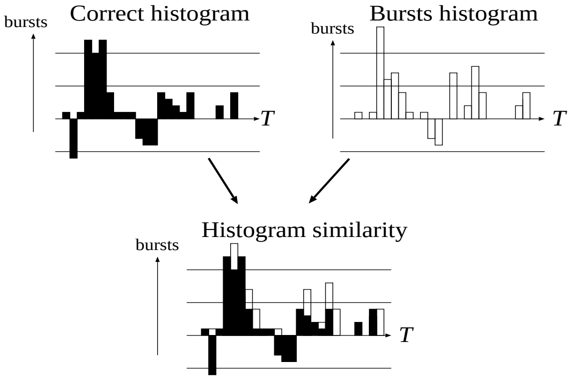

Figure 2: Topic detection with MACD

1. Create document-based MACD histogram where X-axis refers to T, *i.e.*, a period of time (numbered from day 1 to 365). Y-axis is the document count in A per day. Hereafter, referred to as correct histogram.

2. Create term-based MACD histogram where X-axis refers to T, and Y-axis denotes bursts of word w in A. Hereafter, referred to as bursts histogram.

3. We assume that if a term w is informative for summarizing a particular documents in a collection, its burstiness approximates the burstiness of documents in the collection. Because w is a representative word of each document in the task. Based on this assumption, we computed similarity between correct and word histograms by using KL-distance[2]. Let P and Q be a normalized distance of correct histogram, and bursts histogram, respectively. KL-distance is defined by $D(P \mid\mid Q) = \sum_{i=1} P(x_i) \log \frac{P(x_i)}{Q(x_i)}$ where x_i refers bursts in time i. If the value of $D(P \mid\mid Q)$ is smaller than a certain threshold value, w is regarded as a topic word.

3 Extrinsic Evaluation to Summarization

3.1 Event detection

An event word is something that occurs at a specific place and time associated with some specific actions (Allan, 2003; Allan et al., 1998). It refers to notions of who(person), where(place),

[1] In the experiment, we set n_1, n_2, and n_3 to 4, 8 and 5, respectively (He and Parker, 2010).

[2] We tested KL-distance, histogram intersection and Bhattacharyya distance to obtain similarities. We reported only the result obtained by KL-distance as it was the best results among them.

when(time) including what, why and how in a document. Therefore, we can assume that named entities(NE) are linguistic features for event detection. An event word refers to the *theme* of the document itself, and frequently appears in the document but not frequently appear in other documents. Therefore, we first applied NE recognition to the target documents to be summarized, and then calculated tf$*$idf to the results of NE recognition. We extracted words whose tf$*$idf values are larger than a certain threshold value, and regarded these as event words.

3.2 Sentence extraction

We recall that our hypothesis about key sentences in multiple documents is that they include topic and event words. Each sentence in the documents is represented using a vector of frequency weighted words that can be event or topic words.

Like much previous work on extractive summarization (Erkan and Radev, 2004; Mihalcea and Tarau, 2005; Wan and Yang, 2008), we used Markov Random Walk (MRW) model to compute the rank scores for the sentences. Given a set of documents to be summarized, G = (S, E) is a graph reflecting the relationships between two sentences. S is a set of vertices, and each vertex s_i in S is a sentence. E is a set of edges, and each edge e_{ij} in E is associated with an affinity weight $f(i \rightarrow j)$ between sentences s_i and s_j ($i \neq j$). The affinity weight is computed using cosine measure between the two sentences, s_i and s_j. Two vertices are connected if their affinity weight is larger than 0 and we let $f(i \rightarrow i)$= 0 to avoid self transition. The transition probability from s_i to s_j is then defined as follows:

$$p(i \rightarrow j) \; = \; \begin{cases} \frac{f(i \rightarrow j)}{\sum_{k=1}^{|S|} f(i \rightarrow k)}, & \text{if } \Sigma f \; \neq \; 0 \\ 0 \,, & \text{otherwise.} \end{cases} \qquad (7)$$

We used the row-normalized matrix U_{ij} = $(U_{ij})_{|S| \times |S|}$ to describe G with each entry corresponding to the transition probability, where U_{ij} = $p(i \rightarrow j)$. To make U a stochastic matrix, the rows with all zero elements are replaced by a smoothing vector with all elements set to $\frac{1}{|S|}$. The final transition matrix is given by formula (8), and each score of the sentence is obtained by the principal eigenvector of the matrix M.

$$M \;\; = \;\; \mu U^T + \frac{(1-\mu)}{\mid S \mid} \vec{e}\vec{e}^T. \qquad (8)$$

We selected a certain number of sentences according to rank score into the summary.

4 Experiments

4.1 Experimental settings

We applied the results of topic detection to extractive multi-document summarization task, and examined how the results of topic detection affect the overall performance of the salient sentence selection. We used two tasks, Japanese and English summarization tasks, NTCIR-3[3] SUMM Japanese and DUC[4] English data. The baselines are (i) MRW model (**MRW**): The method applies the MRW model only to the sentences consisted of noun words, (ii) Event detection (**Event**): The method applies the MRW model to the result of event detection, (iii) Topic Detection by LDA (**LDA**): MRW is applied to the result of topic candidates detection by LDA and (iv) Topic Detection by LDA and MACD (**LDA & MACD**): MRW is applied to the result of topic detection by LDA and MACD only, *i.e.*, the method does not include event detection.

4.2 NTCIR data

The data used in the NTCIR-3 multi-document summarization task is selected from 1998 to 1999 of Mainichi Japanese Newspaper documents. The gold standard data provided to human judges consists of FBFREE DryRun and FormalRun. Each data consists of 30 tasks. There are two types of correct summary according to the character length, "long" and "short", All series of documents were tagged by CaboCha (Kudo and Matsumoto, 2003). We used person name, organization, place and proper name extracted from NE recognition (Kudo and Matsumoto, 2003) for event detection, and noun words including named entities for topic detection. FBFREE DryRun data is used to tuning parameters, *i.e.*, the number of extracted words according to the tf$*$idf value, and the threshold value of KL-distance. The size that optimized the average Rouge-1(R-1) score across 30 tasks was chosen. As a result, we set tf$*$idf and KL-distance to 100 and 0.104, respectively.

We used FormalRun as a test data, and another set consisted of 218,724 documents from 1998 to 1999 of Mainichi newspaper as a corpus used in

[3] http://research.nii.ac.jp/ntcir/

[4] http://duc.nist.gov/pubs.html

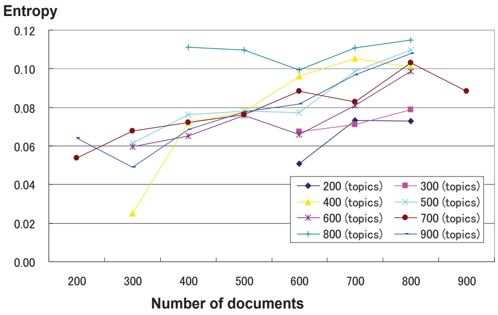

Figure 3: Entropy against the # of topics and documents

Method	Short	Long
	R-1	R-1
MRW	.369	.454
Event	.625	.724
LDA	.525	.712
LDA & MACD	.630	.742
Event & Topic	.678	.744

Table 1: Sentence Extraction (NTCIR-3 test data)

LDA and MACD. We estimated the number of k' and d' in LDA, *i.e.*, we searched k' and d' in steps of 100 from 200 to 900. Figure 3 illustrates entropy value against the number of topics k' and documents d' using 30 tasks of FormalRun data. Each plot shows that at least one of the documents for each summarization task is included in the cluster. We can see from Figure 3 that the value of entropy depends on the number of documents rather than the number of topics. From the result shown in Figure 3, the minimum entropy value was 0.025 and the number of topics and documents were 400 and 300, respectively. We used them in the experiment. The summarization results are shown in Table 1.

Table 1 shows that our approach, "Event & Topic" outperforms other baselines, regardless of the summary type (long/short). Topic candidates include surplus words that are not related to the topic because the results obtained by "LDA" were worse than those obtained by "LDA & MACD", and even worse than "Event" in both short and long summary. This shows that integration of LDA and MACD is effective for topic detection.

4.3 DUC data

We used DUC2005 consisted of 50 tasks for training, and 50 tasks of DUC2006 data for testing in order to estimate parameters. We set tf∗idf and KL-distance to 80 and 0.9. The minimum entropy value was 0.050 and the number of topics and documents were 500 and 600, respectively. 45 tasks from DUC2007 were used to evaluate the performance of the method. All documents were tagged by Tree Tagger (Schmid, 1995) and Stanford Named Entity Tagger [5] (Finkel et al., 2005). We used person name, organization and location for event detection, and noun words including named entities for topic detection. AQUAINT corpus[6] which consists of 1,033,461 documents are used as a corpus in LDA and MACD. Table 2 shows Rouge-1 against unigrams.

Method	R-1	Method	R-1
MRW	.381	Event	.407
LDA	.402	LDA & MACD	.428
Event & Topic	**.438**		
PYTHY	.426	HybHSum	**.456**
hPAM	.412	TTM	**.447**

Table 2: Comparative results (DUC2007 test data)

We can see from Table 2 that Rouge-1 obtained by our approach was also the best compared to the baselines. Table 2 also shows the performance of other research sites reported by (Celikylmaz and Hakkani-Tur, 2010). The top site was "HybHSum" by (Celikylmaz and Hakkani-Tur, 2010). However, the method is a semi-supervised technique that needs a tagged training data. Our approach achieves performance approaching the top-performing unsupervised method, "TTM" (Celikylmaz and Hakkani-Tur, 2011), and is competitive to "PYTHY" (Toutanoval et al., 2007) and "hPAM" (Li and McCallum, 2006). Prior work including "TTM" has demonstrated the usefulness of semantic concepts for extracting salient sentences. For future work, we should be able to obtain further advantages in efficacy in our topic detection and summarization approach by disambiguating topic senses.

5 Conclusion

The research described in this paper explores a method for detecting topic words over time in series of documents. The results of extrinsic evaluation showed that integration of LDA and MACD is effective for topic detection.

[5] http://nlp.stanford.edu/software/CRF-NER.shtml

[6] http://catalog.ldc.upenn.edu/LDC2002T31

References

J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. 1998. Topic Detection and Tracking Pilot Study Final Report. In *Proc. of the DARPA Broadcast News Transcription and Understanding Workshop*.

J. Allan, editor. 2003. *Topic Detection and Tracking*. Kluwer Academic Publishers.

D. M. Blei and J. D. Lafferty. 2006. Dynamic Topic Models. In *Proc. of the 23rd International Conference on Machine Learning*, pages 113–120.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet Allocation. In *The Journal of Machine Learning Research*, volume 3, pages 993–1022.

A. Celikylmaz and D. Hakkani-Tur. 2010. A Hybird Hierarchical Model for Multi-Document Summarization. In *Proc. of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 815–824.

A. Celikylmaz and D. Hakkani-Tur. 2011. Discovery of Topically Coherent Sentences for Extractive Summarization. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 491–499.

G. Erkan and D. Radev. 2004. LexPageRank: Prestige in Multi-Document Text Summarization. In *Proc. of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 365–371.

J. R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 363–370.

G. Folino, C. Pizzuti, and G. Spezzano. 2007. An Adaptive Distributed Ensemble Approach to Mine Concept-Drifting Data Streams. In *Proc. of the 19th IEEE International Conference on Tools with Artificial Intelligence*, pages 183–188.

F. Fukumoto, Y. Suzuki, A. Takasu, and S. Matsuyoshi. 2013. Multi-document summarization based on event and topic detection. In *Proc. of the 6th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 117–121.

D. He and D. S. Parker. 2010. Topic Dynamics: An Alternative Model of Bursts in Streams of Topics. In *Proc. of the 16th ACM Special Interest Group on Knowledge Discovery and Data Mining*, pages 443–452.

R. Klinkenberg and T. Joachims. 2000. Detecting Concept Drift with Support Vector Machines. In *Proc. of the 17th International Conference on Machine Learning*, pages 487–494.

R. Klinkenberg. 2004. Learning Drifting Concepts: Example Selection vs. Example Weighting. *Intelleginet Data Analysis*, 8(3):281–300.

T. Kudo and Y. Matsumoto. 2003. Fast methods for kernel-based text analysis. In *Proc. of 41st Annual Meeting of the Association for Computational Linguistics*, pages 24–31.

M. M. Lazarescu, S. Venkatesh, and H. H. Bui. 2004. Using Multiple Windows to Track Concept Drift. *Intelligent Data Analysis*, 8(1):29–59.

W. Li and A. McCallum. 2006. Pachinko Allocation: Dag-Structure Mixture Model of Topic Correlations. In *Proc. of the 23rd International Conference on Machine Learning*, pages 577–584.

K. Mane and K. Borner. 2004. Mapping Topics and Topic Bursts in PNAS. *Proc. of the National Academy of Sciences of the United States of America*, 101:5287–5290.

R. Mihalcea and P. Tarau. 2005. Language Independent Extractive Summarization. In *In Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 49–52.

J. Murphy. 1999. *Technical Analysis of the Financial Markets*. Prentice Hall.

L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The Pagerank Citation Ranking: Bringing Order to the Web. In *Technical report, Stanford Digital Libraries*.

H. Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proc. of the European chapter of the Association for Computational Linguistics SIGDAT Workshop*.

M. Scholz. 2007. Boosting Classifiers for Drifting Concepts. *Intelligent Data Analysis*, 11(1):3–28.

R. Swan and J. Allan. 2000. Automatic Generation of Overview Timelines. In *Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 38–45.

K. Toutanoval, C. Brockett, M. Gammon, J. Jagarlamudi, H. Suzuki, and L. Vanderwende. 2007. The Phthy Summarization System: Microsoft Research at DUC. In *Proc. of Document Understanding Conference 2007*.

X. Wan and J. Yang. 2008. Multi-Document Summarization using Cluster-based Link Analysis. In *Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306.