社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－DBS－143 (69)
2007／7／3

ユーザから指定された時刻に焦点を当てる文書クラスタリング法

キーソァポァン† 石川 佳治†† 北川 博之†,†††

† 筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻 〒305-8573 茨城県つくば市天王台1-1-1

†† 名古屋大学情報連携基盤センター 〒464-8601 愛知県名古屋市千種区不老町

††† 筑波大学計算科学研究センター 〒305-8573 茨城県つくば市天王台1-1-1

E-mail: †sophoin@kde.cs.tsukuba.ac.jp, ††ishikawa@itc.nagoya-u.ac.jp, †††kitagawa@cs.tsukuba.ac.jp

あらまし 大量の電子化された文書の出現により，文書データのマイニングに関する研究が盛んとなっている．本研究では，ユーザが指定した興味ある時刻の周辺において，過去のイベントを要約するための文書クラスタリング法を提案する．過去・未来の双方向に指数的に逓減する関数を導入することで，着目する時点の文書には1を，過去・未来の文書については着目時点から離れるほど小さい重みを割り当て，クラスタリングに反映する．提案手法の性能をTDT2データセットに基づいて評価する．

キーワード K-means, 指数的逓減，確率的類似度，双方向の逓減ファクター，ユーザの興味

A Document Clustering Method Focusing on User Specified Time of Interest

Sophoin KHY†, Yoshiharu ISHIKAWA††, and Hiroyuki KITAGAWA†,†††

† Department of Computer Science, Graduate School of Systems and Information Engineering, University of Tsukuba, Tennoudai 1-1-1, Tsukuba, Ibaraki, 305-8573 Japan

†† Information Technology Center, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601 Japan

††† Center for Computational Sciences, University of Tsukuba, Tennoudai 1-1-1, Tsukuba, Ibaraki, 305-8573 Japan

E-mail: †sophoin@kde.cs.tsukuba.ac.jp, ††ishikawa@itc.nagoya-u.ac.jp, †††kitagawa@cs.tsukuba.ac.jp

Abstract With the explosion of immense amount of electronic documents, intensive research to mine important information has accordingly emerged. This paper presents a document clustering method which summarizes past events around a user specified time of interest. Incorporating an exponential two-sided decay function, the method assigns the highest weights or values *one* to documents at the focused time and smaller and smaller weights to documents which precede or succeed the focused time. The performance of the method is investigated through experimental evaluation on TDT2 data set.

Key words K-means, exponential decay, probabilistic similarity, two-sided decay factor, user's interest

1. Introduction

With the explosion of immense amount of electronic documents, intensive research of varying objectives to mine important information has accordingly emerged. Document clustering, which organizes a set of objects such that similar objects are grouped into the same clusters while dissimilar ones are grouped into different clusters, has been used as a core technique in managing vast amount of data and providing summarized information [1], [3].

In addition, events happen somewhere in the world everyday. The events could be, for instance, meetings between leaders of two countries, natural disasters like tsunamis, accidents like airplane crashes, or news about celebrities, etc. As a matter of fact, they gradually fade away from our memory as time passed. It is therefore necessary to trace back to past events to recall or to learn more about it. In this paper, we address this problem of obtaining past events

from chronologically ordered stored data. Specifically, our research concerns a document clustering approach focusing on users' interest on past events. The study can address a user's question like "what has happened around January 1st this year?".

The objective of this research is to provide a summary of past events happened around a specific period of time in the past specified by the user. Specifically, given the user specified time of interest, the method generates topics around the specified time. Moreover, it 'discriminates' documents whose timestamps precede or succeed the user specified time. In this paper, an exponential two-sided decay function is proposed and incorporated into the similarity measure. By the incorporation of such a decay function, the approach assigns the highest weights *one* to documents acquired on the user specified time of interest and exponentially degrading weights from the user's interest time to documents whose timestamps precede or succeed the focus time.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed approach where the exponential two-sided decay function, similarity measure incorporating the decay function and the clustering algorithm are described. Section 4 reports the experimental evaluation on a subset of TDT2 Corpus. Section 5 concludes the paper.

2. Related work

There has been extensive work in document clustering. Conventional document clustering methods, in general, focus on developing efficient and effective clustering methods, which group similar documents into a same cluster. Jain et al. provides a general survey of clustering method in [5]. Although the approach proposed in this paper employs a document clustering method, the approach is not only a clustering approach. Chronological order of documents according to their timestamps (or arrival time), decay function and user interaction are incorporated in the clustering method in this paper.

Additionally, in the retrospective topic detection task in the TDT competition [1], [11] which discovers previously unidentified topics in a chronologically ordered accumulation of documents, a number of clustering methods have been proposed. Although chronological order of documents [9] and timestamps [7] have been employed in the research, the incorporation of decay function and user specification of interest as in the approach in this paper were not considered in the research.

3. Proposed approach

In this section, the exponential two-sided decay function is described. The incorporation of the decay function into the similarity measure and the clustering algorithm come next.

3.1 A Two-sided decay function

A two-sided decay function to represent the decay of a document's value from the user's specified time of interest (or focal point) is proposed in this approach. It is based on the idea that user interest affects importance of documents. It causes decrease in the importance of documents as they appear far from the user specified focus time of interest.

[Definition 1] (Document weight) *Let the user's specified time of interest be $t = T$ and the acquisition time of each document d_i be T_i, or simply timestamp of the document. We define the* weight *of a document d_i, dw_i, by*

- If $T_i \leq T$,

$$dw_i \stackrel{\text{def}}{=} \lambda_1^{T-T_i}. \quad (1)$$

- If $T_i > T$,

$$dw_i \stackrel{\text{def}}{=} \lambda_2^{T_i-T}. \quad (2)$$

In the above definition, λ_1, λ_2, $(0 < \lambda_1, \lambda_2 < 1)$, are called *decay factors*. They are parameters tuned according to the target document set.

The parameters λ_1 and λ_2 are user specified parameters. To help users to decide the value of the parameter, a metaphor to impart intuitive meanings is recommended. Users are suggested to gives a *half-life span* β value which specifies the period that a document loses half of its importance. Namely, β satisfies $\lambda^\beta = 1/2$. The forgetting factor λ then can be derived as follows:

$$\lambda = \exp\left(-\frac{\log 2}{\beta}\right) \quad (3)$$

Figure 1 depicts the two-sided exponential decay of document weights. Weights of documents at the user specified time T are *one*, the highest, while weights of documents before and after T decrease exponentially towards both sides according to the rates specified by the decay factors λ_1 (on the left side of T) and λ_2 (on the right side of T).

3.2 Similarity measure

The exponential two-sided decay function described in the above section is incorporated into the similarity measure proposed in [4]. In this section, the similarity measure is briefly described.

In the following, documents in the document set are represented by d_i $(i = 1, \ldots, n)$ and all index terms in the document set by t_k $(k = 1, \ldots, m)$.

The *subjective probability* $\Pr(d_i)$ to randomly select a document d_i from the document set is defined as follows:

$$\Pr(d_i) \stackrel{\text{def}}{=} \frac{dw_i}{tdw}, \quad (4)$$

where dw_i is the weight of d_i shown in Eq. (1) and (2), and

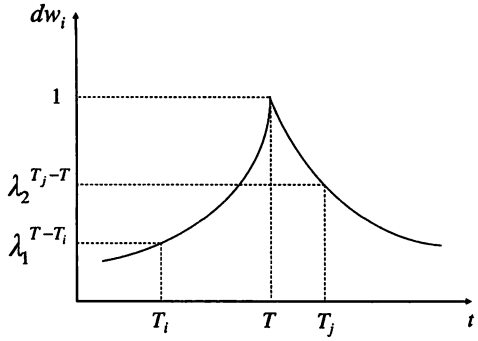

図 1 Two-sided decay function

tdw is the total weight of all documents in the document set:

$$tdw \stackrel{\text{def}}{=} \sum_{i=1}^{n} dw_i. \tag{5}$$

The selection probability of a document is proportional to its weight. This means that 'interested' documents have larger selection probabilities than 'uninterested' ones.

Next the conditional probability $\Pr(t_k|d_i)$ that term t_k is selected from document d_i is derived. The probability is simply derived based on the number of occurrences of terms in a document.

$$\Pr(t_k|d_i) \stackrel{\text{def}}{=} \frac{f_{ik}}{\sum_{l=1}^{m} f_{il}}, \tag{6}$$

where t_k is an index term and f_{ik} is the number of occurrences of term t_k within document d_i.

The occurrence probability of t_k in the entire document set can be derived by

$$\Pr(t_k) = \sum_{i=1}^{n} \Pr(t_k|d_i) \cdot \Pr(d_i). \tag{7}$$

Using the above formulas and the Bayes' theorem,

$$\Pr(d_j|t_k) = \frac{\Pr(t_k|d_j)\Pr(d_j)}{\Pr(t_k)}. \tag{8}$$

is obtained.

Then, the conditional probability $\Pr(d_j|d_i)$ can be expanded as

$$\Pr(d_j|d_i) = \sum_{k=1}^{m} \Pr(d_j|d_i, t_k)\Pr(t_k|d_i). \tag{9}$$

It is generally assumed that $\Pr(d_j|d_i, t_k) \simeq \Pr(d_j|t_k)$ is approximately hold, then

$$\Pr(d_j|d_i) \simeq \sum_{k=1}^{m} \Pr(d_j|t_k)\Pr(t_k|d_i). \tag{10}$$

Based on the above formulas,

$$\Pr(d_i, d_j) = \Pr(d_j|d_i) \cdot \Pr(d_i) \tag{11}$$

$$\simeq \frac{\Pr(d_i)\Pr(d_j)}{\sum_{l=1}^{m} f_{il} \sum_{l=1}^{m} f_{jl}} \sum_{k=1}^{m} \frac{f_{ik} f_{jk}}{\Pr(t_k)}. \tag{12}$$

This formula says that the co-occurrence probability between the two documents is based on their importance to the user, basically implied by $\Pr(d_i)$ and $\Pr(d_j)$, and the contents of the documents.

Next, the above formulas is transformed to a more simple one using vector representation.

The document vector $\boldsymbol{d_i}$ of d_i is defined as

$$\boldsymbol{d_i} \stackrel{\text{def}}{=} (tf_{i1} \cdot idf_1, tf_{i2} \cdot idf_2, \ldots, tf_{im} \cdot idf_m), \tag{13}$$

where tf_{ik} is the *term frequency* of t_k within d_i

$$tf_{ik} \stackrel{\text{def}}{=} f_{ik}, \tag{14}$$

and idf_k is the *inverse document frequency (IDF)* of t_k

$$idf_k \stackrel{\text{def}}{=} \frac{1}{\sqrt{\Pr(t_k)}}. \tag{15}$$

Let len_i be the *document length* of d_i:

$$len_i \stackrel{\text{def}}{=} \sum_{l=1}^{m} f_{il}. \tag{16}$$

Using the vector representation, Eq. (12) can be transformed as:

$$\Pr(d_i, d_j) = \Pr(d_i)\Pr(d_j)\frac{\boldsymbol{d_i} \cdot \boldsymbol{d_j}}{len_i \times len_j}. \tag{17}$$

This co-occurrence probability $\Pr(d_i, d_j)$ is derived based on the notion of document novelty and the $tf \cdot idf$ weighting scheme of the conventional vector space model.

Finally, the similarity metric is defined as follows:

$$sim(d_i, d_j) \stackrel{\text{def}}{=} \Pr(d_i, d_j). \tag{18}$$

This definition says the co-occurrence probability between d_i and d_j is used for their similarity score. $\Pr(d_i, d_j)$ is a probability to select d_i and d_j when we select two documents randomly from the document repository. The probability will be large when two documents have similar term occurrence patterns and they appear near the user specified focus time of interest. On the other hand, the probability will be small when two documents do not share same terms and/or at least one of the documents appears far from the focus time.

3.3 Clustering algorithm

This section introduces the extended K-means clustering algorithm proposed in [6], which is used as the clustering algorithm in the approach in this paper.

- **Initial process**

(1) Select K documents randomly and form initial K clusters.

(2) Compute cluster representatives.

(3) Compute intra-cluster similarities and clustering index G.

- **Iteration process**

(1) For each document d, do the following two steps:

(a) For each cluster, compute the intra-cluster similarity when d is appended to the cluster.

(b) Assign d to a cluster such that the assignment causes the largest increase in the intra-cluster similarity. If no assignment increases the intra-cluster similarity, put d into the outlier list.

(2) Recompute cluster representatives.

(3) Recompute G and take it as G_{new} .

(4) If $(G_{\text{new}} - G_{\text{old}})/G_{\text{old}} < \delta$, terminate, where δ is a pre-defined constant.

(5) Otherwise, return to Step 1.

The clustering index G introduced above is a measure used to control the convergence criterion of the clustering. At each iteration, the index is used to evaluate the quality of the clustering and to decide whether to stop the clustering process. It is defined as follows.

$$G \stackrel{\text{def}}{=} \sum_{p=1}^{K} |C_p| \cdot avg\_sim(C_p), \tag{19}$$

where K is the number of clusters to be generated, $|C_p|$ is the number of documents in cluster C_p and $avg\_sim(C_p)$ is the *average similarity* of documents in cluster C_p and is defined as:

$$avg\_sim(C_p) \stackrel{\text{def}}{=} \frac{1}{|C_p|(|C_p|-1)} \sum_{d_i \in C_p} \sum_{d_j \in C_p,\, d_i \neq d_j} sim(d_i, d_j). \tag{20}$$

$avg\_sim(C_p)$ is the *intra-cluster similarity* and is used as a measure to decide the goodness and poorness of a clustering result.

4. Experiments

This section describes the experimental methodology to evaluate the performance of the method and presents the results.

4.1 Data set

The experimental data is a subset of the TDT2 Corpus [10] which has been used in TDT (Topic Detection and Tracking) competitions [11], consisting of news articles associated with topic labels. The data contains 7,578 documents corresponding to 96 topics, dated January 4th, 1998 to June 30th, 1998.

In this experiment, a portion of the subset of the TDT2 Corpus, Jan 4th to Feb 1st, is used. The statistic of the data is given below.

- No. of docs = 1,735
- No. of topics = 30
- Min. topic size = 1
- Max. topic size = 453
- Med. topic size = 14.5
- Mean topic size = 273

Some topics contained in the Jan4-Feb1 data are presented in Table 1.

表 1 Topic of TDT2 corpus from Jan4-Feb1 1998

Topic ID	Count	Topic Name
20001	453	Asian Economic Crisis
20002	276	Monica Lewinsky Case
20008	19	Casey Martin Sues PGA
20012	140	Pope visits Cuba
20013	76	1998 Winter Olympics
20015	337	Current Conflict with Iraq
20023	66	Violence in Algeria
20031	21	John Glenn
20033	79	Superbowl '98
20077	93	Unabomber

4.2 Evaluation framework

The experimental framework is prepared as follows:

- Data set: Jan04-Feb01 (29 days), where we assume Jan04-Feb01 is the user specified time interval.
- User specified point in time of interest (focal point): Jan 18
- Parameter setting:
  - $\beta_1 = \beta_2 = 4$ ($\lambda = 0.84$), k = 10
  - $\beta_1 = \beta_2 = 7$ ($\lambda = 0.91$), k = 16
  - $\beta_1 = \beta_2 = 14$ ($\lambda = 0.95$), k = 16

Clustering results are evaluated by the following performance measures in IR research [2], [8]:

$$\text{Precision:} \quad p = \frac{a}{a+b} \tag{21}$$

$$\text{Recall:} \quad r = \frac{a}{a+c} \tag{22}$$

$$F = \frac{2rp}{r+p} \tag{23}$$

where a is the number of documents corresponding to a topic in a cluster generated by a system; $a+b$ is the total number of documents in a cluster generated by a system; $a+c$ is the total number of documents in a topic given by the evaluation data. F is the harmonic mean of recall and precision.

For each cluster, the precision, recall and F measures are computed. A cluster is *assigned* a topic if the precision of the topic in the cluster is greater than a predefined threshold. In the experiments, we use 0.50 as the threshold value. If a cluster does not have such a topic, the cluster is not marked with any topic.

4.3 Results

In this section, the experimental results are presented.

Figures 2 through 4 show the F scores of the clustering results using $\beta = 4$, $\beta = 7$ and $\beta = 14$, respectively. Additionally, a summary of the F scores of the three clustering results with processing time is presented in Table 2. The F score for $\beta = 4$ is the smallest one while the F score for $\beta = 14$ is the largest. This is arguably a result of different decay rates driven by the β values. For $\beta = 4$, the decay factor is $\lambda = 0.84$ which is undoubtedly a fast decay rate, while in the case of $\beta = 14$, $\lambda = 0.95$. As a result, small decay factors cause documents unable to take part in the clustering and thus result in low performance scores. Consequently, this, however, causes gathering of documents around the focal point only.

	$\beta = 4$	$\beta = 7$	$\beta = 14$
No. of cluster	9	12	11
Average F score	0.49	0.59	0.73
Processing time	8min9sec	10min31sec	11min4sec

表 2 Overview of the results

Figures 5 through 7 show histograms of topic 20012 ("Pope Visits Cuba") of the clustering results (grey color and upside down) ploted against the TDT2 evaluation data (dark green color). Similarly, Figures 8 through 10 present histograms of topic 20015 ("Current Conflict with Iraq") of the clustering results (grey color and upside down) ploted against the TDT2 evaluation data (dark green color). In the same way, histograms of topic 20077 ("Unabomber") of the clustering results (grey color and upside down) ploted against the TDT2 evaluation data (dark green color) are depicted in Figures 11 through 13.

In each figure of the three topics, the result for $\beta = 4$ are the most focused or densed around the focal point, the user specified time of interest Jan 18, than other β's, thanks to the small β which results in fast decay of document weights. A larger β, $\beta = 7$, slightly expands from $\beta = 4$, while $\beta = 14$ stretchs towards both ends from the focal point, owing to the gentle degradation of decay of document weights. Setting of a small β value is conclusively more suited to the context of the approach in this paper since results obtained contain documents focused around the focal point than a large value of β one does.

5. Conclusion

Managing documents in large scale and providing useful information is an important issue. This paper introduced an approach for clustering documents focusing on user's interest. The approach proposed an incorporation of the two-sided decay function into a clustering method and user indication of time of interest such that importance of documents vary depending on their position related to the focal point in order to provide clusters of topics around the focal point. Different clustering results were obtained as shown in the experiments owing to the introduction of such a decay function to the clustering method. This approach has important application in document management area like news summarization.

Other decay functions should also be considered other than the two-sided decay function. While weblog has rapidly broadened in size and gained wide attention from research community, the extension of the research to this area should be an interesting research issue.

Acknowledgements

This research is partly supported by the Grant-in-Aid for Scientific Research (19300027) from Japan Society for the Promotion of Science (JSPS) and the Grant-in-Aid for Scientific Research (19024006) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. In addition, this work is supported by the grants from Hoso Bunka Foundation.

文　献

[1] Allan, J. (ed.): Topic Detection and Tracking: Event-based Information Organization. Kluwer, Boston (2002)

[2] Baeza-Yates, R., and Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Harlow, England (1999)

[3] Chen, H., Kuo, J., Huang, S., Lin, C., Wung, H.: A Summarization System for Chinese News from Multiple Sources. Journal of the Ameri. Socie. for Info. Sci. and Tech. (JASIST), **54**(13), pp. 1224–1236 (2003)

[4] Ishikawa, Y., Chen, Y., Kitagawa, H.: An On-line Document Clustering Method Based on Forgetting Factors. Proc. of 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), Darmstadt, Germany, September 4-9, pp. 325–339, (2001)

[5] Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Computing Surveys **31**(3), (1999)

[6] Khy, S., Ishikawa, Y., Kitagawa, H.: A novelty-based clustering method for on-line documents, World Wide Web Journal, DOI 10.1007/s11280-007-0018-9

[7] Li, Z., Wang, B., Li, M., Ma, W-Y.: A probabilistic model for retrospective news event detection. In: Proc. 28th ACM SIGIR Conference, pp. 106–113 (2005)

[8] van Rijsbergen, C.J.: Information Retrieval. Butter Worths, Sydney (1979)

[9] Yang, Y., Carbonell, J.G., Brown, R.G., Pierce, T., Archibald, B.T., Liu, X.: Learning Approaches for Detecting and Tracking News Event. IEEE Intel. Sys. **14**(4), July/August, pp. 32–43 (1999)

[10] Linguistic Data Consortium (LDC), http://www.ldc.upenn.edu/

[11] National Institute of Standards and Technology (NIST), http://www.nist.gov/speech/tests/tdt/

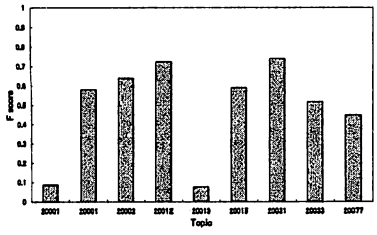

図 2 $\beta = 4$

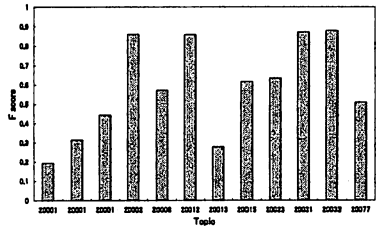

図 3 $\beta = 7$

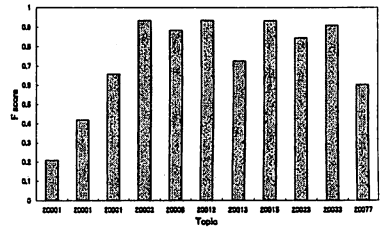

図 4 $\beta = 14$

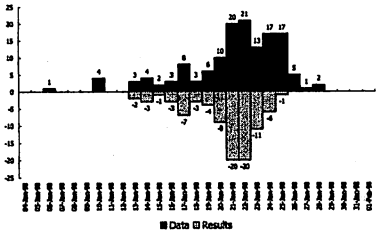

図 5 Topic 20012 with $\beta = 4$

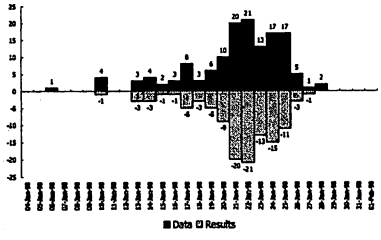

図 6 Topic 20012 with $\beta = 7$

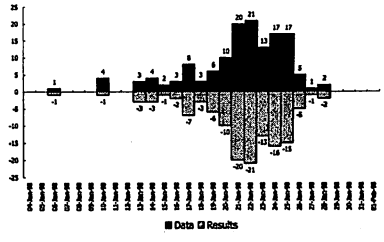

図 7 Topic 20012 with $\beta = 14$

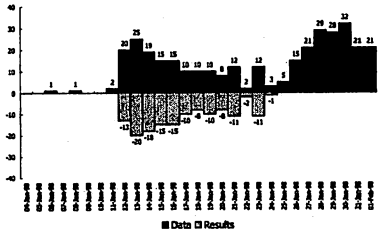

図 8 Topic 20015 with $\beta = 4$

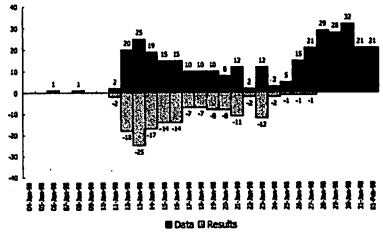

図 9 Topic 20015 with $\beta = 7$

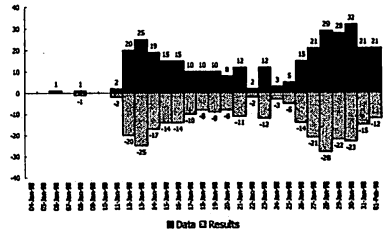

図 10 Topic 20015 with $\beta = 14$

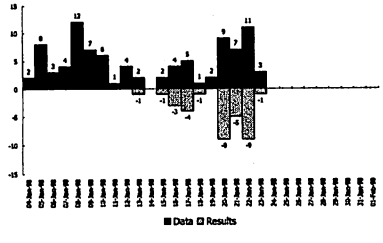

図 11 Topic 20077 with $\beta = 4$

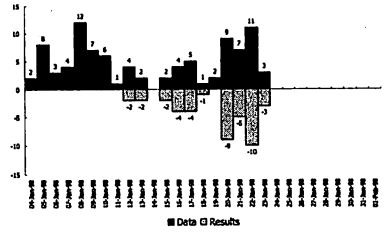

図 12 Topic 20077 with $\beta = 7$

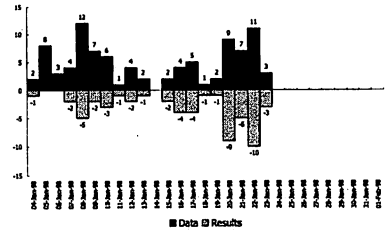

図 13 Topic 20077 with $\beta = 14$