Finding "Coordinate" Relationships in Search Results

Meng ZHAO†, Hiroaki OHSHIMA†, and Katsumi TANAKA†

† Department of Social Informatics, Graduate School of Informatics, Kyoto University
Yoshida Honmachi, Kyoto, 606–8501 Japan
E-mail: †{zhao,ohshima,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract We propose a method to structure search results of a user-given query depending on some perspectives, while these perspectives are detected and extracted from the given data collection, and moreover, confirmed by user perceptions on the Web. For example, in the case of news articles, take as an example the search results of a keyword query "school shooting". They can be grouped by different events, such as "Oregon school shooting", or "Virginia Tech shooting", or by developmental stages of events, such as "occurrence of an event", "number of casualties", or "end of the shooter". However, since it is intellectually easier to understand "school shooting" by considering similar but different events, our method finally chooses to cluster the collection into different events, viz. the former grouping. We aim at providing a more comprehensible way for users to have a whole and exhaustive understanding of unstructured search results by transforming to perspective-based structured ones.

Key words coordinate relationship, clustering, Web search results

1. Introduction

Nowadays, search engines, such as Google(注1) and Bing(注2) play a more and more important role in people's daily life. People can obtain a variety of information they want through search engines. Take two different search services for example. One is "news search", which provides an important access to Web versions of newspapers, magazines, and news wires. Suppose one wants to know recently happened school shootings, and correspondingly, issues a keyword query "school shooting". Figure 1 shows search results returned by Google News(注3) (at the time of writing the document). From it, we can see that similar news articles, such as ones about plan to fight gun violence, or Facebook shooting massacre threat, are aggregated together. However, the presentation of news articles is a simple list, which means it mixes together different events about school shooting. Although users can scan the list from top to bottom until they have found the information they are looking for, it is hard to pick up a certain event, for example, "Newtown school shooting". Moreover, it is also hard to survey the whole topic "school shooting". The same happens when querying Yahoo News(注4). On the other hand, Bing News(注5) cannot find any results for "school shooting". Another example is "academic paper search", which provides a way for users to access academic papers indexed on the Web. Suppose one wants to survey papers on ranking algorithm, and correspondingly, issues a keyword query "ranking algorithm". Figure 2(a) shows papers found by Google Scholar(注6), while Figure 2(b) shows papers found by Microsoft Academic Search(注7). From the figures, we can find that similar situations happen. The presentation of papers is a simple list. They do not show any connections or relations between papers. For example, paper A is a rival of paper B. Therefore, it is hard to survey a research area by scanning the paper list. From the above discussion, we can find that the major problem is the unstructured search results.

Many previous work have tackled the problem by apply clustering method [16] [17] [11] [20]. Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Therefore, by clustering search results, similar Web pages (news articles or academic papers in the above examples) are classified into the same group. However, results from the state-of-the-art clustering methods might deviate from human cognition. In the previous "school shooting" example, the following two news articles, titled "Oregon School Shooting: 1 Student Dead, Suspect Also Killed" and "Second Suspect Pleads Guilty in Frederick High School Shoot-

(注1)：http://www.google.com
(注2)：http://www.bing.com
(注3)：https://news.google.com
(注4)：http://news.yahoo.com
(注5)：https://www.bing.com/news
(注6)：https://scholar.google.com
(注7)：http://academic.research.microsoft.com

ing", respectively, are more likely to be grouped in the same cluster, since both of them describe the injuries and deaths during shootings. However, for human beings, it is natural and intuitive to capture these two news articles as different events. One is "Oregon School Shooting", and another is "Frederick High School Shooting".

Oregon School Shooting: 1 Student Dead, Suspect Also Killed

At a press conference following the Oregon school shooting this morning, police say one student was gunned down at Reynolds High School in Troutdale, Ore. According to USA Today, officials later confirmed that the shooter was also deceased. At this point...

Second Suspect Pleads Guilty in Frederick High School Shooting

The second of two men charged in a shooting that wounded two teenage boys outside Frederick High School at a basketball game pleaded guilty for his role in the incident Tuesday. Brandon Earl Tyler, 22, pleaded guilty to two counts of first-degree assault.

Our objective is to structure search results of a user-given query in such a way that it is intellectually easier for users to understand the aggregation result. In order to accomplish this goal, we introduce "coordinate relationship" between pairs of documents. In brief, two documents are coordinate to each other if they talk about the same topic, and moreover, their primary actions are similar, and doers or receivers of actions are correspondingly coordinate. Our notion is that coordinate documents should not be grouped into the same cluster, since they are mutually exclusive in semantics. Hence, we first detect coordinate documents in the Web search results, and then modify a clustering algorithm by giving a penalty to the distance between paris of coordinate documents.

The remainder of the paper is organized as follows. We dedicate Section 2. to the discussion of the previous work on classic clustering, event detection and tracking in news stream, and scientific literature clustering. We then discuss the key concept of this work, "coordinate relationship" in Section 3.. In Section 4., we describe the details of our proposed method. In Section 5., we show experimental results on news search and give an analysis of them. Finally, we conclude the paper in Section 6..

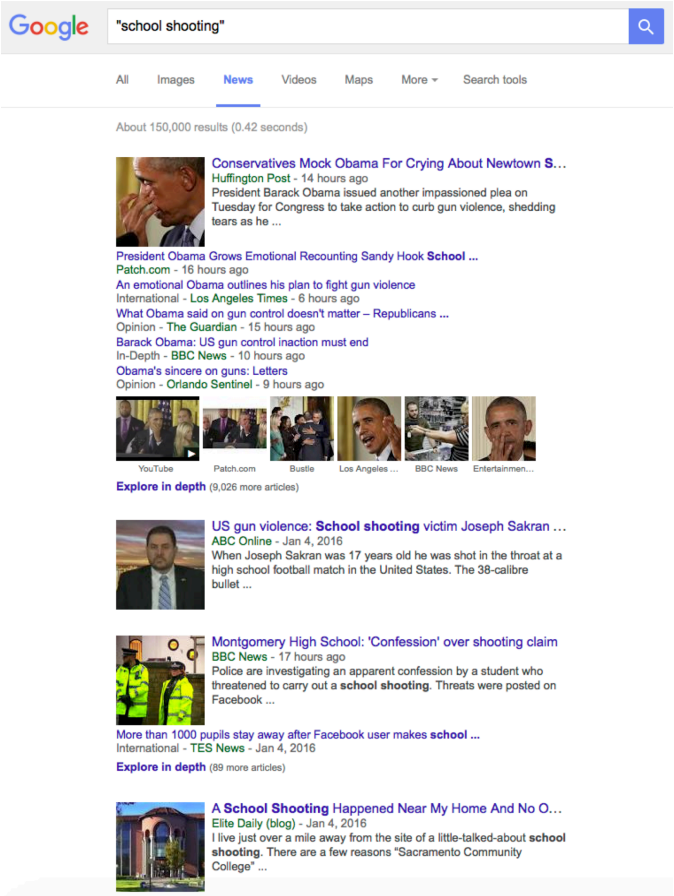


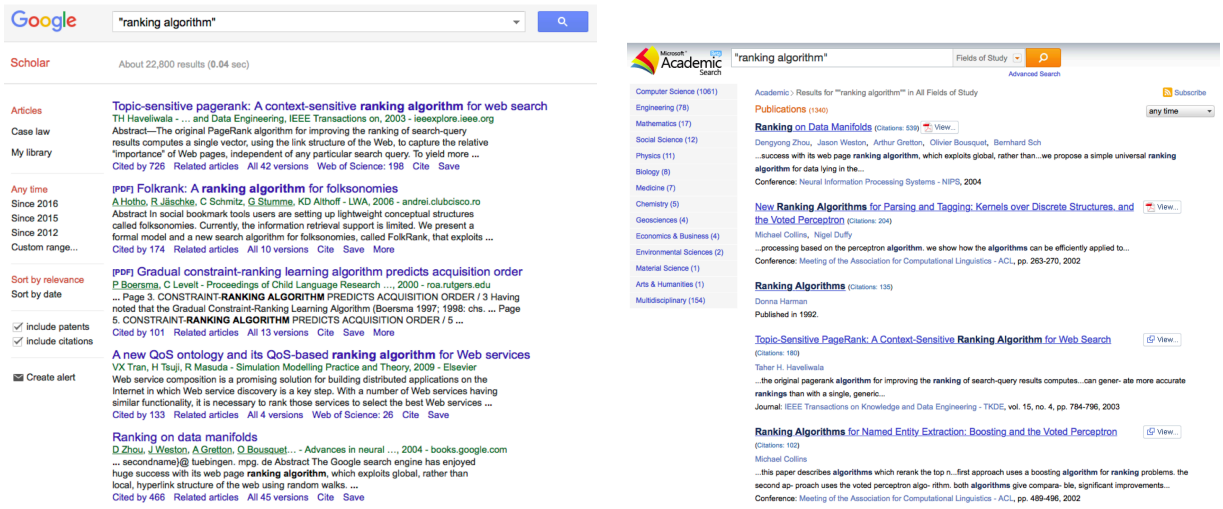
Figure 1 Search results of "school shooting" by Google News.

2. Related Work

2.1 Clustering

Clustering algorithms aim to create groups (called clusters) that are coherent internally, but clearly different from each other. In other words, objects within the same cluster should be as similar as possible; and objects in one cluster should be as dissimilar as possible from objects in other clusters.

According to the structure of clusters, they can be broadly grouped into two categories: flat clustering and hierarchical clustering. Flat clustering [3] [10] creates a flat set of clusters, which means that its clusters are independent of each other. K-means [10] is the most important flat clustering algorithm, which takes an iterative refinement technique. Given K, the number of clusters, the first thing to do is to randomly select K objects as initial cluster centers. The algorithm then moves the cluster centers around in space in order to minimize the average squared Euclidean distance of objects from their cluster centers. This is done iteratively by alternating between two steps until a stopping criterion is met: reassigning objects to the cluster with the closest centroid, and recalculating each centroid based on the objects



(a) (b)

Figure 2 Search results of "ranking algorithm".

in each new cluster. On the other hand, hierarchical clustering [6] [7] [15] results in a hierarchy of clusters, which means its clusters can be visualized using a tree structure (a dendrogram). Its algorithms are either top-down (divisive) or bottom-up (agglomerative). For example, bottom-up algorithms regard each object as a singleton cluster at the very beginning and then agglomerate pairs of clusters until all clusters have been merged into a single cluster that contains all objects. Besides, hierarchical clustering algorithms do not need a pre-given number of clusters.

2.2 News Event Mining

Allan et al. introduced the concept of new event detection (NED) in [1]. NED requires identifying news stories that discuss an event that has not already been reported in the past. They proposed a possible definition of event as something that happens at a particular time and place. However, there are some problems with this definition. In their discussion, it is hard to define event, but it is easier to define parts of event identity, the properties that make two events the same. Therefore, part of their problem became deciding what properties of news stories can be used to distinguish different events. They found that news stories about the same event often occur in clumps and there must be something about the story that makes its appearance worthwhile. Finally, they used a single pass clustering algorithm and a thresholding model that incorporates the properties of events to detect new events. Kumaran and Allan [8] tackled the same problem and employed text classification techniques as well as named entities to improve the performance of NED.

Compared with NED, Yang et al. proposed the concept of retrospective news event detection (RED) in [19]. RED is defined as the discovery of previously unidentified events in historical news corpus. Both of contents and time information contained in a news article are very helpful for RED. Since many work, including [19], only focus on the utilization of the contents, Li et al. [9] considered better representations of news events, which effectively models both the contents and time information.

Nallapati et al. [13] made a similar attempt as us. They also considered that viewing a news topic as a flat collection of stories is not efficient for users to understand the topic quickly. Hence, they introduced an event model to capture the rich structure of news events and their dependencies in a news topic, such as the causality or temporal-ordering between pairs of news events. Their algorithm first groups news stories into unique events in the topic by using the agglomerative clustering with time decay, and then constructs dependencies among them.

Feng and Allan [4] [5] also dealt with the problem that news topics are treated as a flat list, ignoring the intrinsic connection among each stories. In [4], they clustered text passages and then created links with scenario-specific rules to generate incident threading. While in [5], they removed the assumption that a news story covers a single topic, and consequently, extended the incident threading to passage level.

3. Coordinate Relationship

Here, we introduce a key concept in this paper, called "coordinate relationship".

"Coordinate relationship" exists in different levels, such as term level, sentence level, passage level and document level. In term level, two terms are coordinate to each other if a term shares a hypernym with another. For example, since both "lemon" and "grapefruit" belong to citrus fruit category, they are coordinate terms. Here, "citrus fruit" is their common hypernym. When it comes to sentence level,

we make an extension as stated in our previous work [21]. We believe a sentence can be mapped by a template and an entity tuple. Take as an example the sentence "lemons are rich in vitamin c". It can be generated by the template ***X*** *are rich in* ***Y*** and the entity tuple *(lemons, vitamin c)*. Another entity tuple, such as *(apples, pectin)*, is coordinate to *(lemons, vitamin c)*, since there also exists the highConcentration relation between "apples" and "pectin" such that apples contain a high amount of pectin. As a consequence, two sentences are coordinate to each other if they satisfy the following two conditions: (1) their templates convey the same meaning; (2) their entity tuples are coordinate to each other. In the aforementioned example, the sentence "lemons are rich in vitamin c" is coordinate to the sentence "apples contain a high amount of pectin".

Let us go one step further and discuss "coordinate relationship" in document level. It can be regarded as an extension of that in sentence level. To understand the "coordinate relationship" in document level, we present an example. Consider the two news articles given below.

Article 1: Oregon shooting

The Oregon shooting ***occurred*** *on October 1, 2015 at the UCC campus near Roseburg, Oregon, United States. Christopher Harper-Mercer, a 26-year-old enrolled at the school, fatally* ***shot*** *an assistant professor and eight students in a classroom. Seven to nine others were* ***injured****. After being wounded by two police officers, the gunman* ***committed suicide*** *by shooting himself in the head.*

Article 2: Virginia Tech shooting

The Virginia Tech shooting ***occurred*** *on April 16, 2007, on the campus of Virginia Polytechnic Institute and State University in Blacksburg, Virginia, United States. Seung-Hui Cho, a senior at Virginia Tech,* ***shot and killed*** *32 people and* ***wounded*** *17 others in two separate attacks, approximately two hours apart, before* ***committing suicide****.*

While both these articles talk about school shootings and their casualties, they obviously are about different events. In human cognition, it is easy to distinguish them as two different events under the topic "school shooting". We found that there are two main reasons. First, the primary actions in both these articles are similar (see terms or phrases in bold). For example, in the beginning of both these articles, it states the occurrence of an event, using exactly the same term "occurred". When it comes to the injuries and deaths of each shootings, article 1 used the term "injured", while article 2 used the term "wounded". Though these two terms have different surface forms, they have the same semantic meaning. Second, the doers of similar actions are somekind of coordinate to each other. So are the receivers or the affected of similar actions, such as the time and place where events took place. In more details, the occurrence time of the Oregon shooting, *October 1, 2015*, is coordinate to the occurrence time of the Virginia Tech shooting, *April 16, 2007*. Similarly, the occurrence location of the Oregon shooting, *UCC campus*, is coordinate to that of the Virginia Tech shooting, *campus of Virginia Polytechnic Institute and State University*.

Based on the above discussion, we know that primary actions in documents can be regarded as an extension of templates in sentences, but in a more abstract level; doers or receivers of actions form an extension of entity tuples in sentences, no longer limited to two entities. As a consequence, in the following, we focus on these two observations to detect coordinate documents.

4. Our Method

As we mentioned in Section 1., we believe that the presentation of search results in a flat list is difficult for users to understand the intrinsic connection among individual search result. Therefore, our attempt is to group search results into clusters such that documents in the same cluster talk about the same event; and documents in different clusters describe different events under the same topic (also called "coordinate events"). Besides, within each cluster, we also show the development of an event, such as the beginning, the outcome, or the impact of an event.

Our notion is that coordinate documents should not be grouped into the same cluster, since they are mutually exclusive in semantics. Hence, we first detect coordinate documents in the Web search results, and then modify a clustering algorithm by giving a penalty to the distance between paris of coordinate documents.

4.1 Finding Coordinate Documents

As we discussed in Section 3., two documents are coordinate to each other if they talk about the same topic, and moreover, their primary actions are similar, and doers or receivers of actions are correspondingly coordinate. Since we devote to help users quickly and better capture search results of their queries, we assume that the issued query is a *topic* that they concern with. Consequently, search results of a certain query are regarded to talk about the same topic. As a result, the problem of finding coordinate documents turns to the following two sub-problems:

- confirming coordinate relationships between pairs of doers or receivers.
- confirming semantic similarities between pairs of actions.

Since doers or receivers are nouns or noun phrases, and actions are described by verbs, we extract nouns/noun phrases

and verbs and treat them in different ways.

4.1.1 Coordinate Relationship Confirmation

Since two terms are coordinate to each other if they share a common hypernym, it is intuitive to verify whether two terms are coordinate to each other by finding if they share any common hypernyms. Therefore, an intuitive way is to check dictionaries. WordNet [12] is a lexical database for the English language, widely used as a dictionary or thesaurus. In WordNet, terms are grouped into sets of synonyms called synsets, while each synset expresses a distinct concept. Moreover, synsets are interlinked by means of conceptual-semantic and lexical relations. Given a synset, more general synsets (hypernyms) can be found. As a result, we can confirm coordinate relationship between a pair of terms (or part of phrases) by using hypernym information in WordNet.

However, just like other dictionaries, for terms or phrases not in WordNet, nothing can be found through it. So we turn our eyes on the Web, the greatest data factory. But we do not find shared hypernyms anymore. Instead, we directly check whether two terms are coordinate or not by coordinate conjunctions, such as "and", or "or". The idea is from a bi-directional lexico-syntactic pattern-based algorithm [14], which can quickly acquire coordinate terms (or phrases) for any query term (phrase). Suppose we want to confirm the coordinate relationship between *Umpqua Community College* and *Virginia Polytechnic Institute and State University*. The following two queries are generated and queried on the Web.

- "*Umpqua Community College* **and** *Virginia Polytechnic Institute and State University*"
- "*Virginia Polytechnic Institute and State University* **and** *Umpqua Community College*"

If both these two queries can have search results returned, it shows that *Umpqua Community College* is coordinate to *Virginia Polytechnic Institute and State University*. Otherwise, they are not coordinate to each other.

Besides, since we focus on certain topics, the queries given by users, we also take topics into consideration. It means when we check the coordinate relationship, we also consider the context where terms (or phrases) are. Therefore, in the "school shooting" example, the aforementioned two queries turn to be

- "*Umpqua Community College* **and** *Virginia Polytechnic Institute and State University*" AND(注8) "school shooting"
- "*Virginia Polytechnic Institute and State University* **and** *Umpqua Community College*" AND "school shooting"

Then, similarly, the confirmation is based on their occurrence on the Web.

(注8)：Notice that this "and" indicates an AND search rather than a conjunction in natural language.

Algorithm 1 Constrained Hierarchical Agglomerative Clustering Algorithm

Input: Set of documents $D = \{d_1, d_2, ..., d_N\}$,
set of pairs of coordinate documents $\mathbb{C} = \{(d_i, d_j)\}$,
distance function $dist(d_i, d_j)$.
Output: Disjoint partitioning $C = \{C_1, C_2, ...\}$of D.

```
for i = 1 to N do
  for j = 1 to N do
    I[i][[j] ⇐ dist(d_i, d_j)
  end for
end for
C ⇐ []
for k = 1 to N − 1 do
  < i, m > ⇐ arg min_{<i,m>:i≠m} I[i][m]
  C.APPEND(< i, m >)
  for n = 1 to N do
    I[i][j] ⇐ dist(i, m, n)
    I[j][i] ⇐ dist(i, m, n)
  end for
end for
```

4.1.2 Semantic Similarity Confirmation

Since we can get synonyms of terms from WordNet, a simple way to check whether two verbs are semantically similar is to see whether they are synonyms in WordNet. However, still we have the problem that nothing can be found if the verb is not in WordNet. We again use the Web to confirm semantic similarity between verbs. In our previous work [21], we stated that if two templates convey the same meaning, they share many common entity tuples. Similarly, if two verbs are semantically similar to each other, they should have many common entity tuples. Based on this hypothesis, we extract entity tuples for each verb from the Web and regard them as the context of each verb. Suppose we want to check whether "wound" is a synonym of "hurt". Since in most cases, search results of a single verb cover vast quantities of different information, we use the user given query for limitation. Still take the "school shooting" for example. For "wound", the generated query is *wound AND "school shooting"*, while for "hurt", the generated query is *hurt AND "school shooting"*. Nouns or Noun phrases just before or just after the verb are extract from the top N Web search results. They construct entity tuples, further. For example, we have *(gunshot, jaw)*, *(gunshot, head)*, *(shot, arm)* for "wound". Then these entity tuples are used to generate a vector, which represents the context of a certain verb. A tf/idf weighting is applied to weigh each element in vectors. Two verbs are similar if the cosine similarity of their vectors are greater than a threshold θ.

4.1.3 Constrained HAC based on Coordinate Documents

As we discussed before that coordinate documents are mutually exclusive, they should not be grouped in the same cluster. It is a meaningful selection of **cannot-link** constraints(注9) [18]. In previous work [2] [18], cannot-link pairs are selected by manual, while we aim at choosing these constraints automatically.

Our clustering algorithm is shown in Algorithm 1. We first compute the $N \times N$ distance matrix I, where penalties are given based on coordinate degree between documents. The algorithm then executes $N - 1$ steps of merging the currently most similar clusters. In each iteration, the two most similar clusters are merged and correspondingly, there is a update in I. The clustering is stored as a list of merges in C. The function $dist(i, m, n)$ computes the distance between cluster C_n and the new cluster that merged cluster C_i and C_m. Single-link, complete-link or average-link can be used here.

5. Evaluations

Currently, we concentrate on News search. In other words, the search results we tackle with are news articles at present. Since our objective is to structure search results of a user-given query, we plan to prepare 100 keyword queries for evaluation. Besides, those 100 queries are chosen from different categories, such as world, business, technology, entertainment, sports, science and health. Our claim is that with the help of our organization of search results, it is intellectually easier for users to understand a topic. Hence, a user study will be constructed to demonstrate that our structure of search results better fits human cognition.

6. Conclusion

The presentation of search results in a flat list is difficult for users to understand the intrinsic connection among individual search result. Besides, it is also hard to survey a user-given topic by scanning the flat list. As a result, we proposed a method to structure search results of a user-given query in such a way that it is intellectually easier for users to understand the aggregation result. We defined two documents are coordinate to each other if they talk about the same topic, and moreover, their primary actions are similar, and doers or receivers of actions are correspondingly coordinate. Our basic idea is that coordinate documents should not be grouped into the same cluster, since they are mutually exclusive in semantics. Hence, we first detect coordinate documents in the Web search results, and then modify a clustering algorithm by giving a penalty to the distance between paris of coordinate documents.

(注9)：Cannot-link constraints specify that two in- stances must not be placed in the same cluster.

Acknowledgment

This work was supported in part by the following projects: Grants-in-Aid for Scientific Research (Nos. 15H01718 and 24680008) from MEXT of Japan.

References

[1] Allan, J., Papka, R., Lavrenko, V.: On-line new event detection and tracking. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 37–45. SIGIR '98, ACM (1998)

[2] Basu, S., Banerjee, A., Mooney, R.J.: Active semi-supervision for pairwise constrained clustering. In: Proceedings of the 2004 SIAM International Conference on Data Mining (SDM-04) (April 2004)

[3] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, Series B 39(1), 1–38 (1977)

[4] Feng, A., Allan, J.: Finding and linking incidents in news. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management. pp. 821–830. CIKM '07, ACM (2007)

[5] Feng, A., Allan, J.: Incident threading for news passages. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. pp. 1307–1316. CIKM '09, ACM (2009)

[6] F.Murtagh: A survey of recent advances in hierarchical clustering algorithms. The Computer Journal 26, 354–359 (1983)

[7] G.N.Lance, W.T.Williams: A general theory of classificatory sorting strategies. The Computer Journal 9, 373–380 (1967)

[8] Kumaran, G., Allan, J.: Text classification and named entities for new event detection. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 297–304. SIGIR '04, ACM (2004)

[9] Li, Z., Wang, B., Li, M., Ma, W.Y.: A probabilistic model for retrospective news event detection. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 106–113. SIGIR '05, ACM (2005)

[10] MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. pp. 281–297 (1967)

[11] Mecca, G., Raunich, S., Pappalardo, A.: A new algorithm for clustering search results. Data Knowl. Eng. 62(3), 504–522 (Sep 2007)

[12] Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM 38(11), 39–41 (1995)

[13] Nallapati, R., Feng, A., Peng, F., Allan, J.: Event threading within news topics. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. pp. 446–453. CIKM '04, ACM (2004)

[14] Ohshima, H., Oyama, S., Tanaka, K.: Searching coordinate

terms with their context from the web. In: Proceedings of WISE. pp. 40–47 (2006)
[15] Olson, C.F.: Parallel algorithms for hierarchical clustering. Parallel Computing 21(8), 1313–1325 (1995)
[16] Osinski, S., Stefanowski, J., Weiss, D.: Lingo: Search results clustering algorithm based on singular value decomposition. In: Proceedings of the International Intelligent Information Processing and Web Mining. pp. 359–368. IIPWM '04, Springer Berlin Heidelberg (2004)
[17] Osinski, S., Weiss, D.: A concept-driven algorithm for clustering search results. IEEE Intelligent Systems 20(3), 48–54 (May 2005)
[18] Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S.: Constrained k-means clustering with background knowledge. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 577–584. ICML '01, Morgan Kaufmann Publishers Inc. (2001)
[19] Yang, Y., Pierce, T., Carbonell, J.: A study of retrospective and on-line event detection. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 28–36. SIGIR '98, ACM (1998)
[20] Zeng, H.J., He, Q.C., Chen, Z., Ma, W.Y., Ma, J.: Learning to cluster web search results. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 210–217. SIGIR '04, ACM (2004)
[21] Zhao, M., Ohshima, H., Tanaka, K.: Sentential query rewriting via mutual reinforcement of paraphrase-coordinate relationships. In: Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services. iiWAS 2015 (2015)