News Contextualization with Geographic and Visual Information

Zechao Li\textsuperscript{1,2}, Meng Wang\textsuperscript{1}, Jing Liu\textsuperscript{1}, Changsheng Xu\textsuperscript{1,2} and Hanqing Lu\textsuperscript{1,2}
\textsuperscript{1}Institute of Automation, Chinese Academy of Sciences, Beijing China 100190
\textsuperscript{2}China-Singapore Institute of Digital Media, 21 Heng Mui Keng Terrace, Singapore 119613
\textsuperscript{3}School of Computing, National University of Singapore, Singapore
\{zcli, jliu, csxu, luhq\}@nlpr.ia.ac.cn, eric.mengwang@gmail.com

ABSTRACT
In this paper, we investigate the contextualization of news documents with geographic and visual information. We propose a matrix factorization approach to analyze the location relevance for each news document. We also propose a method to enrich the document with a set of web images. For location relevance analysis, we first perform toponym extraction and expansion to obtain a toponym list from news documents. We then propose a matrix factorization method to estimate the location-document relevance scores while simultaneously capturing the correlation of locations and documents. For image enrichment, we propose a method to generate multiple queries from each news document for image search and then employ an intelligent fusion approach to collect a set of images from the search results. Based on the location relevance analysis and image enrichment, we introduce a news browsing system named NewsMap which can support users in reading news via browsing a map and retrieving news with location queries. The news documents with the corresponding enriched images are presented to help users quickly get information. Extensive experiments demonstrate the effectiveness of our approaches.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; H.5.1 [Information Interfaces and Presentation]: Multimedia Information System

General Terms
Algorithms, Experimentation

Keywords
News location relevance, Image enrichment, NewsMap

\*Area chair: Cees Snoek

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MM’11, November 28–December 1, 2011, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-4503-0616-4/11/11 ...$10.00.

1. INTRODUCTION

With the proliferation of news documents on the Internet, online news reading becomes an important approach for information acquisition in people’s daily lives. People can access and browse news on either some main web portals, such as MSN and Yahoo!, or large news websites, such as CNN, AOL and MSNBC. It is reported by Newspapers Association of America (NAA) that, in US, there are more than 100 millions of unique visitors on several major news websites monthly and the number is nearly 2/3 of all adult Internet users\textsuperscript{1}. However, existing news presentation usually has the following two limitations:

(1) Lack of location-based organization. Existing studies demonstrate that users usually have high priority in getting news information about some specific places, such as his/her country, working city and hometown [29]. Most large news websites can organize news documents according to the involved countries, such that a rough personalization can be accomplished by identifying the user’s location based on IP address. Of course we can search for news documents by giving a location name as query on many news websites, but the location names contained in news documents are usually noisy and this will degrade the search performance.

(2) Incomprehensive visual information. A picture is worth a thousand words. The associated pictures, as the complement of news text, are able to provide readers additional information or help them get information quickly. However, the pictures contained in news documents are usually very few. In fact, a statistical analysis on a randomly collected dataset (see the description about the dataset in Section 6) shows that more than half of the news documents do not contain any picture and only less than 5% of the news documents contain more than one picture.

To address these problems, we propose approaches to contextualize news documents with geographic and visual information respectively. We employ a matrix factorization approach to analyze the location relevance for each news document, and enrich the document with a set of web images. It is well known that the key part of a news document is “4W” elements: “Who”, “When”, “Where” and “What” [22]. Our approaches are actually able to enhance the three of them, namely, “Who”, “Where” and “What”. The enhancement

\textsuperscript{1}http://www.naa.org/TrendsandNumbers.aspx
What elements in news documents (see (a) and (b), respectively).

of “Where” is obvious as we can estimate the location relevance for the document. The other elements are enhanced by the image enrichment. For example, as shown in Fig. 1, when reading a news document about a new movie of Julia Roberts, the user may be interested in who this actor is. When reading a news document about an ice hockey competition, the user may be interested in what the sport is like. By providing several illustrative images, our approach is able to provide the user with more vivid and comprehensive information that complements the text document.

Our approaches work as follows. We employ a toponym extraction and expansion process to extract the location names from news documents. However, they are usually noisy and incomplete. For example, there are many documents that contain multiple locations or do not contain any location. The relevance of locations can be inferred by mining the content of news documents and their correlations. For example, a news document discussing about “Olympic 2008” may not contain a location name, but we can infer its relevance to “Beijing” by exploring other topic-related but location-indicated news documents. To address this problem, we propose a probabilistic matrix factorization algorithm, called Correlation Consistent Probabilistic Matrix Factorization (CCPMF), to analyze the location relevance of documents. For image enrichment, a set of web images is collected with the help of Google image search. The title and main body of each document are analyzed to generate queries with different lengths for image search, and we then employ an intelligent fusion method to merge the search results. The pictures contained in news documents will also be explored in the image enrichment process.

Based on the news location analysis and image enrichment approaches, we introduce a news browsing system named NewsMap. Users can browse the map and click a place, and the system will show the news documents that are relevant to the place. It also supports news search with a textual location query. The enriched images are presented together with the title and summary for each news document such that users can quickly get necessary information. We employ a PageRank approach that incorporates both the location relevance and the time information to rank news documents.

Our contribution can be summarized as follows:

1) We propose a matrix factorization based algorithm to analyze the location relevance for news documents. In comparison with the conventional matrix factorization approach, our method investigates the news content correlation and location co-occurrence and thus achieves much better performance.

2) We propose an approach to enrich news documents with a set of web images. A set of techniques, such as query generation, rank aggregation and duplicate detection are integrated to find appropriate images. Although it involves several existing techniques, such as rank aggregation and duplicate detection, our approach that generates queries with different lengths and then fuses their search results is novel, and it can also be applied in other applications.

3) Based on the location relevance analysis and image enrichment, we introduce a news browsing system named NewsMap, which supports users in browsing news documents via a map.

The rest of this paper is organized as follows. Section 2 reviews related work. In Section 3 and Section 4, we introduce the location relevance estimation and image enrichment approaches for news documents, respectively. We introduce the NewsMap system in Section 5. Experimental results are reported in Section 6. We conclude the paper with future work in Section 7.

2. RELATED WORK

Since our work involves news location analysis and enriching text documents with images, we mainly review previous work along these two directions.

2.1 Media Geolocation

Media geolocation has attracted great research interest in recent years for its many potential applications, such as media search, recommendation and travel assistance.

Hays and Efros [13] proposed an image geolocation approach which locates an image by finding its nearest neighbors in a GPS-tagged dataset. Serdyukov et al. [28] proposed a text-driven method to analyze the geographic information of Flickr images. A language model based on tags is learned for each grid of map and the location of an image is determined by these models. Existing work has shown that the performance of web image geolocation can be improved by combining textual and visual features [3, 7]. The literature regarding video geolocation is relatively sparse. Kelm et al. [19] proposed an approach that combines textual and visual models to find the grid of a video. Christel et al. [4] proposed a method to tag video sequences with latitude and longitude information. It supports users in browsing videos via a map with all of the places discussed in the videos. However, it cannot handle noisy and incomplete place names. While both providing map interface for browsing, our NewsMap system provides a snap view that illustrates the news titles and the related images in order to help readers get information quickly (details will be introduced in Section 5). More details about image and video geolocation can be found in [23, 39]. Different from the geolocation of image and video that needs to explore visual content and tag information, our task is to analyze the relevance levels of the location names contained in news documents. Therefore, in this work we employ a matrix factorization approach with exploring the correlation of news documents and locations.
The task of finding a geographic focus of a web page was first proposed in [10]. Two methods are investigated, one is by analyzing the geographic location of the hosts linking into the site and the other is based on the detection and disambiguation of location name. The methods in [1, 41] rely on propagating the confidence weights of locations up to the root of the gazetteer taxonomy to determine the place each name refers to. But these methods cannot be applied to our task as we do not have links among news documents. Co-occurrence models are used to disambiguate place names in [25]. In this method, Wikipedia is used to generate a co-occurrence model, which can be used to improve average precision in the geographic information retrieval system. However, the task of [25] is place name disambiguation, while our target is to analyze the location relevance of documents and it is able to handle the noisy and incomplete locations for documents. Geographical Information Retrieval (GIR), which is the augmentation of Information Retrieval with geographic metadata, is also related to our work. To provide the necessary framework for evaluating GIR systems, a track named GeoCLEF [12] started from 2005. GIR tries to exploit geographical references in text for targeted and improved retrieval. Different from the tasks of GeoCLEF and GIR, the target of our work is to analyze the location relevance of documents and support users to search news by place names as queries.

Currently, there are several systems that support news browsing with geographic information, such as World News Map\(^2\) and Yahoo News Map\(^3\). However, the geolocation is only performed at country level based on several simple heuristics. The NewsMap system introduced in our work can support news browsing with locations at different scales and can provide much better user experience.

### 2.2 Text Visualization

There exists work that aims to find images to describe or enrich text documents. WordsEye [6] employed semantic representation to synthesize 3D scenes from text. Joshi et al. [18, 40] proposed a story picturing system that is able to find a set of pictures to describe a fragment of a text. In their approach, key phrases are extracted from the text to retrieve images, which are then ranked using lexical annotations and visual content. In [40], the picturability is measured in terms of the ratio of the frequencies under regular web versus image search. Graphics is used to spatially arrange the images to represent the story. Several methods are also proposed to find or synthesize an image to describe a web page [16, 30]. However, there is little work on enriching news documents with media information. The most related work is [8] and [9]. Delgado et al. [8, 9] proposed methods to assist users in reading news by providing images. However, the methods select images from a fixed database based on their text information, such as tags, whereas our approach applies image search tools to find images from the web. In addition, the methods by Delgado et al. add images for each sentence, whereas our work finds several images to enrich the whole news document together with the original pictures. In fact, the scheme by Delgado et al. is more likely designed for tools that assist children and elderly, whereas our approach can enhance context information for general users.

### 3. LOCATION RELEVANCE ANALYSIS

In this section, we elaborate our news location analysis approach, which consists of two steps: (1) toponym extraction and expansion and (2) matrix factorization based relevance analysis.

#### 3.1 Toponym Extraction and Expansion

We obtain a toponym list from the news data set with toponym extraction and expansion approach based on external knowledge resources. The process is described as follows. First, we employ OpenNLP\(^4\), a homogeneous package based on a machine learning approach using maximum entropy, to extract toponym candidates from news documents. Second, we remove several candidates that do not have geo-coordinates in Wikipedia\(^5\). The next step is to expand the toponym list and eliminate geographical ambiguity. For example, there are two places named “Paris”, one in France and the other in US. We take each toponym candidate as a query and submit it to GeoNames\(^6\), which will return a list of related place names. We collect the corresponding “Name”, “Country”, “Latitude” and “Longitude” information for each returned item. We parse the returned items and de-duplicate the toponym list by exploring the information of “Country” and geo-coordinates. Meanwhile, we employ the information of “Country” and geo-coordinates to eliminate ambiguity among the toponym candidates (note that the information of “Country” contains hierarchical administrative division, such as country and state).

#### 3.2 Matrix Factorization Relevance Analysis

Up to now, we have obtained a toponym list including a set of locations, and we can find the locations contained in each news document. However, a problem is that the locations found in such way are usually noisy. It comes from two facts. First, several location names contained in a news document are not very relevant. For example, considering a news document that describes the wedding of a celebrity, the document may contain some place names about the bride, such as where she was born and the hometown name of the bride, but these locations are not relevant enough in comparison with the location where the wedding is hold. The second fact is that several relevant locations may not appear in the documents. Clearly, to accurately analyze the most relevant location for a news document needs understanding of the news content and it is a challenging NLP problem. Existing studies demonstrate that the semantic space spanned by text keywords can be approximated by a smaller subset of salient words of the original space [38]. News content has the low rank property as well. Therefore, considering a relevance score matrix \( R \) of which the \((i, j)\)-th element indicates the relevance of the \(i\)-th toponym and \(j\)-th document, \( R \) could be approximated by the product of two low-rank matrices \( P \) and \( E \). These two low rank matrices are the latent feature representations of locations and news contents. Thus the relevance analysis problem becomes the task of estimating the location-document relevance matrix with information loss minimization, and we employ matrix factorization to estimate the location-document relevance matrix. We further explore the correlation of locations and news

---

\(^1\)http://incubator.apache.org/opennlp/
\(^2\)http://www.tsmaps.com/
\(^3\)http://isithackday.com/hacks/placemaker/map.php
\(^4\)http://www.geonames.org/
\(^5\)http://www.wikipedia.org/
documents, i.e., the relevance scores of highly correlated locations or similar documents should be close. Therefore, we propose a Correlation ConsistentProbabilistic Matrix Factorization (CCPMF) algorithm to simultaneously integrate the multiple assumptions.

We first introduce some notations. Assume there are \( M \) locations and \( N \) news documents. Let \( \mathbf{R} \in \mathbb{R}^{M \times N} \), \( \mathbf{P} \in \mathbb{R}^{H \times N} \), \( \mathbf{E} \in \mathbb{R}^{H \times N} \), \( \mathbf{C} \in \mathbb{R}^{M \times M} \) and \( \mathbf{S} \in \mathbb{R}^{N \times N} \) denote the location-document relevance matrix, location latent feature matrix, document latent feature matrix, location correlation matrix, and document correlation matrix, respectively. Denote by \( \mathbf{R}^0 \) the location-document appearance matrix, i.e., \( \mathbf{R}^0_{ij} \) if the \( i \)-th location is contained in the \( j \)-th document, and otherwise \( R^0_{ij} = 0 \). The matrices \( \mathbf{P} \) and \( \mathbf{E} \) indicate latent feature representations that we need to estimate in the matrix factorization approach. We usually set \( H < \min(M, N) \) and it is the rank of \( \mathbf{R} \).

To mine the latent relations among the sparse data set, Probabilistic Matrix Factorization (PMF) [27] can be employed which approximates \( \mathbf{R} \) as a product of two lowerrank matrices. However, it does not take the correlations of locations and documents into account, i.e., the facts that the relevance scores of highly correlated locations or similar documents should be close. Therefore, we extend the PMF approach to CCPMF to further consider the correlations of locations and documents.

PMF is a probabilistic linear model with Gaussian observation noise, which is defined as

\[
p(\mathbf{R}^0|\mathbf{P}, \mathbf{E}, \sigma^2_R) = \prod_{i=1}^{M} \prod_{j=1}^{N} \mathcal{N}(\mathbf{R}^0_{ij} | \mathbf{P}^T \mathbf{e}_j, \sigma^2_R \mathbf{I}), \tag{1}
\]

where \( \mathcal{N}(x | \mu, \sigma^2) \) is the Gaussian probability density function with mean \( \mu \) and variance \( \sigma^2 \), and \( \delta_{ij} \) is the indicator function that satisfies \( \delta_{ij} = 1 \) if \( \mathbf{R}^0_{ij} > 0 \), and otherwise \( \delta_{ij} = 0 \).

Zero-mean spherical Gaussian priors are placed on the two lower-rank matrices:

\[
p(\mathbf{P}|\sigma^2_P) = \prod_{i=1}^{M} \mathcal{N}(\mathbf{P}_i | 0, \sigma^2_P \mathbf{I}), \tag{2}
\]

\[
p(\mathbf{E}|\sigma^2_E) = \prod_{j=1}^{N} \mathcal{N}(\mathbf{E}_j | 0, \sigma^2_E \mathbf{I}), \tag{3}
\]

where \( \mathbf{I} \) is the identity matrix.

Through a Bayesian inference, the posterior distribution over \( p(\mathbf{P}, \mathbf{E} | \mathbf{R}^0, \sigma^2_R, \sigma^2_P, \sigma^2_E) \) can be obtained. Maximizing the log-posterior is equal to minimizing the following equation:

\[
\mathcal{L} = \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{N} \delta_{ij}(R^0_{ij} - R_{ij})^2 + \frac{\lambda_P}{2} \text{Tr}[\mathbf{P}^T \mathbf{P}] + \frac{\lambda_E}{2} \text{Tr}[\mathbf{E}^T \mathbf{E}], \tag{4}
\]

where \( \mathbf{R} = \mathbf{P}^T \mathbf{E} \), \( \lambda_P = \sigma^2_P / \sigma^2_R \), \( \lambda_E = \sigma^2_E / \sigma^2_R \) and \( \text{Tr}[] \) denotes the trace operation on a matrix.

To consider the inter-correlations, CCPMF embeds the correlations of locations and documents as two consistency constraints on PMF, which can be described by the graphical model in Fig. 2. CCPMF is formulated to minimize the following equation.

\[
\mathcal{L} = \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{N} \delta_{ij}(R^0_{ij} - R_{ij})^2 + \frac{\lambda_P}{2} \text{Tr}[\mathbf{P}^T \mathbf{P}] + \frac{\lambda_E}{2} \text{Tr}[\mathbf{E}^T \mathbf{E}] \\
+ \frac{\lambda_C}{2} \text{Tr}[\mathbf{R}^T \mathbf{L}^C \mathbf{R}] + \frac{\lambda_S}{2} \text{Tr}[\mathbf{R}^T \mathbf{L}^S \mathbf{R}^T]. \tag{5}
\]

In comparison with the conventional PMF approach, the CCPMF method further incorporates two terms \( F_C(\mathbf{R}) \) and \( F_S(\mathbf{R}) \), which consider the location correlation and document correlation, respectively, and \( \lambda_C \) and \( \lambda_S \) are two positive weighting parameters. The term \( F_C(\mathbf{R}) \) enforces the relevance scores of highly correlated locations to be close while \( F_S(\mathbf{R}) \) enforces the relevance scores of highly similar documents are close. Although several parameters are involved in our formulation, we will observe its effectiveness in the experiments in Section 6.

The term \( F_C(\mathbf{R}) \) is defined as

\[
F_C(\mathbf{R}) = \frac{1}{2} \sum_{k=1}^{M} \sum_{i,j=1}^{N} (R_{ik} - R_{jk})^2 C_{ij} = \text{Tr}[\mathbf{R}^T \mathbf{L}^C \mathbf{R}], \tag{6}
\]

where \( \mathbf{L}^C = \mathbf{D}^C - \mathbf{C} \) is the Laplacian matrix and \( \mathbf{D}^C \) is a diagonal matrix defined as \( \mathbf{D}^C_{ij} = \sum_{m=1}^{M} C_{im} \).

Similar to \( F_C(\mathbf{R}) \), \( F_S(\mathbf{R}) \) can be represented as

\[
F_S(\mathbf{R}) = \frac{1}{2} \sum_{k=1}^{M} \sum_{i,j=1}^{N} (R_{ki} - R_{kj})^2 S_{ij} = \text{Tr}[\mathbf{R}^T \mathbf{L}^S \mathbf{R}^T]. \tag{7}
\]

Based on Eq. 6 and Eq. 7, Eq. 5 can be rewritten as

\[
\mathcal{L} = \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{N} \delta_{ij}(R^0_{ij} - R_{ij})^2 + \frac{\lambda_P}{2} \text{Tr}[\mathbf{P}^T \mathbf{P}] + \frac{\lambda_E}{2} \text{Tr}[\mathbf{E}^T \mathbf{E}] \\
+ \frac{\lambda_C}{2} \text{Tr}[\mathbf{R}^T \mathbf{L}^C \mathbf{R}] + \frac{\lambda_S}{2} \text{Tr}[\mathbf{R}^T \mathbf{L}^S \mathbf{R}^T]. \tag{8}
\]

We adopt gradient descent to solve the above optimization problem over \( \mathbf{P} \) and \( \mathbf{E} \). To implement CCPMF, we need to define \( \mathbf{C} \) and \( \mathbf{S} \), i.e., the correlation matrices for locations and documents.

For the correlation of locations, we adopt Google distance [5]. Let \( d_{ij} \) denote the Google distance between the \( i \)-th and \( j \)-th locations and then \( C_{ij} \) is defined as

\[
C_{ij} = \frac{d_{ij}^2}{\sum_k d_{ik}^2}, \tag{9}
\]

where \( \sigma_{ij} \) is set to the median value of the pairwise distances.

For the correlation of documents, we represent the title, summary and main body of each news document with TF-IDF histograms. Therefore, we can estimate the cosine similarity of the titles, summaries and main bodies of each two
news documents. Denote by $S'$, $S^o$ and $S^b$ the similarity matrices got in this way. We simply fuse them to get $S$, i.e.,

$$S = \alpha S' + \beta S^o + (1 - \alpha - \beta) S^b.$$

(10)

We simply set the parameters $\alpha$ and $\beta$ to 1/3.

It is worth mentioning that different from the location-document appearance matrix $R^d$ which is usually highly sparse, the relevance matrix $R$ may not be sparse. However, we can easily sparsify $R$ by setting the elements that are below a threshold to 0. This can facilitate the inverted indexing structure in large-scale search.

4. IMAGE ENRICHMENT

Our image enrichment approach consists of two steps. The first step is query generation, which forms a set of queries by analyzing both the titles and main bodies of news documents, and the second step is image mining and selection.

4.1 Query Generation

To retrieve images from the external sources, it is necessary to extract queries from news documents. One simple method is to adopt the title as query to search. However, current search engines cannot handle long queries well, and usually very limited results or even no result are returned for complex queries [14, 21]. Here we propose an approach to extract a set of queries.

The first step is performing stemming and removal of punctuation and stop-words for the whole news documents. We then need to identify several query terms from each news document. However, a news document is usually too long, which makes the query term selection difficult. Given the fact that the title of a news document is usually a good summarization of the whole document that is manually constructed by a specialist, and it is reasonable to select the query terms from the title. Considering there are several news documents with fairly long titles, we rank the title terms according to their correlation with the content of main body and only select top terms. Denote by $T = \{\text{term}_1, \text{term}_2, ..., \text{term}_u\}$ and $B = \{\text{term}_1, \text{term}_2, ..., \text{term}_v\}$ the set of terms in the title and main body, respectively. The importance score of $\text{term}_i$ can thus be estimated as

$$\text{score}(\text{term}_i) = \frac{1}{v} \sum_{j=1}^{v} \exp\left(-\frac{d^2(\text{term}_i, \text{term}_j)}{\sigma^2}\right).$$

(11)

Here $d(\text{term}_i, \text{term}_j)$ is the Google distance between $\text{term}_i$ and $\text{term}_j$ and $\sigma$ is the median value of all $d(\text{term}_i, \text{term}_j)$.

We select the top $c$ terms in $T$ with the highest importance scores. In the cases that $u < c$, we simply select $c - u$ terms from $B$ with the highest TF-IDF values to complement $T$.

The images are collected based on the c query terms with the help of online image search tools. However, there is a dilemma here: usually using more query terms for search can obtain more accurate and descriptive images, but it will also get less search results or even no result returned. For example, we consider a query term set \{"Yao Ming", "Olympic", "torch", "representative", "nervous"\}. If we use each single term to search, such as "Yao Ming" or "torch", the resulting images are not descriptive for the news document. If we use the combination of multiple terms such as "Yao Ming Olympic torch" to search, better results can be collected. But if we use all the terms, i.e., "Yao Ming Olympic torch representative nervous", the returned results are limited and are also irrelevant to the whole story. Figure 3 demonstrates the comparison of the top results of different queries.

To address this problem, we propose an approach that first generates all the combinations of the terms as queries for search and then fuse the search results by assigning appropriate weights to the queries that are of different lengths. Clearly, from the $c$ terms, we can generate $L = \sum_{k=1}^{c} \binom{c}{k} = 2^c - 1$ queries with their lengths varying from 1 to $c$.

4.2 Image Mining and Selection

We issue each query on Google image search engine and collect the top $h$ ranked results. Thus our next task is to fuse the $L$ ranking lists, which is a rank aggregation problem. Generally there are two approaches for rank aggregation, namely, score-based fusion and rank-based fusion [35, 37, 24]. Here we adopt the score-based fusion approach. For each image in a ranking list, we can define its relevance score based on its position. However, there is a dilemma here: usually using more query terms for search can obtain more accurate and descriptive images, but it will also get less search results or even no result returned. For example, we consider a query term set \{"Yao Ming", "Olympic", "torch", "representative", "nervous"\}. If we use each single term to search, such as "Yao Ming" or "torch", the resulting images are not descriptive for the news document. If we use the combination of multiple terms such as "Yao Ming Olympic torch" to search, better results can be collected. But if we use all the terms, i.e., "Yao Ming Olympic torch representative nervous", the returned results are limited and are also irrelevant to the whole story. Figure 3 demonstrates the comparison of the top results of different queries.

To address this problem, we propose an approach that first generates all the combinations of the terms as queries for search and then fuse the search results by assigning appropriate weights to the queries that are of different lengths. Clearly, from the $c$ terms, we can generate $L = \sum_{k=1}^{c} \binom{c}{k} = 2^c - 1$ queries with their lengths varying from 1 to $c$.

4.2 Image Mining and Selection

We issue each query on Google image search engine and collect the top $h$ ranked results. Thus our next task is to fuse the $L$ ranking lists, which is a rank aggregation problem. Generally there are two approaches for rank aggregation, namely, score-based fusion and rank-based fusion [35, 37, 24]. Here we adopt the score-based fusion approach. For each image in a ranking list, we can define its relevance score based on its position. However, there is a dilemma here: usually using more query terms for search can obtain more accurate and descriptive images, but it will also get less search results or even no result returned. For example, we consider a query term set \{"Yao Ming", "Olympic", "torch", "representative", "nervous"\}. If we use each single term to search, such as "Yao Ming" or "torch", the resulting images are not descriptive for the news document. If we use the combination of multiple terms such as "Yao Ming Olympic torch" to search, better results can be collected. But if we use all the terms, i.e., "Yao Ming Olympic torch representative nervous", the returned results are limited and are also irrelevant to the whole story. Figure 3 demonstrates the comparison of the top results of different queries.

To address this problem, we propose an approach that first generates all the combinations of the terms as queries for search and then fuse the search results by assigning appropriate weights to the queries that are of different lengths. Clearly, from the $c$ terms, we can generate $L = \sum_{k=1}^{c} \binom{c}{k} = 2^c - 1$ queries with their lengths varying from 1 to $c$.

4.2 Image Mining and Selection

We issue each query on Google image search engine and collect the top $h$ ranked results. Thus our next task is to fuse the $L$ ranking lists, which is a rank aggregation problem. Generally there are two approaches for rank aggregation, namely, score-based fusion and rank-based fusion [35, 37, 24]. Here we adopt the score-based fusion approach. For each image in a ranking list, we can define its relevance score based on its position. However, there is a dilemma here: usually using more query terms for search can obtain more accurate and descriptive images, but it will also get less search results or even no result returned. For example, we consider a query term set \{"Yao Ming", "Olympic", "torch", "representative", "nervous"\}. If we use each single term to search, such as "Yao Ming" or "torch", the resulting images are not descriptive for the news document. If we use the combination of multiple terms such as "Yao Ming Olympic torch" to search, better results can be collected. But if we use all the terms, i.e., "Yao Ming Olympic torch representative nervous", the returned results are limited and are also irrelevant to the whole story. Figure 3 demonstrates the comparison of the top results of different queries.

To address this problem, we propose an approach that first generates all the combinations of the terms as queries for search and then fuse the search results by assigning appropriate weights to the queries that are of different lengths. Clearly, from the $c$ terms, we can generate $L = \sum_{k=1}^{c} \binom{c}{k} = 2^c - 1$ queries with their lengths varying from 1 to $c$.
Figure 4: The schematic illustration of NewsMap.

The fourth component is result ranking and visualization. There are many ranking algorithms [26, 34, 33, 36, 11]. Here we adopt a PageRank [26] approach with considering both location relevance and news timeliness. It will prioritize the following three kinds of news: (1) the news documents that are highly relevant to query location; (2) latest news; and (3) the news documents that have many closely related news reports (they usually indicate important news information that has been intensively reported). They actually indicate three properties for news presentation and search: relevance, timeliness and importance.

The time information is an important aspect for news. We quantize the time stamp of each news document as a number of “YYYYMMDD”. For example, the time “Sep. 12 2010” will be quantized to be 20100912. Denote by date_k the quantized date of the k-th document and we then normalize the numbers with two steps:

$$\text{date}_k = \frac{\text{date}_k - \min_j(\text{date}_j)}{\max_k(\text{date}_k) - \min_j(\text{date}_j)}$$

$$\text{date}_k = \frac{\text{date}_k}{\sum_k \text{date}_k}$$

Through CCPMF, we have obtained a relevance score between the query location and each news document. Denote by score_k the relevance score between the location and the k-th document. We normalize them as follows.

$$\text{score}_k = \frac{\text{score}_k}{\sum_j \text{score}_j}$$

In the PageRank algorithm [26], we set the static ranking score of the k-th document to

$$r_k^0 = \frac{\text{date}_k + \text{score}_k}{2}$$

An iterative process is then implemented by performing

$$r_k^{\text{iter}} = (1 - y)r_k^0 + y \sum_j S_{ij}r_j^{\text{iter}-1}$$

until iter \(\geq 1\) achieves the pre-determined number. Here, \(S_{ij}\) is the similarity between the i-th and j-th document, and y is the damping factor (we simply set it to 0.85 [2]).

A preliminary UI of the NewsMap system is presented in Fig. 5, which is developed based on Google Map API. By submitting a location query (in the top textbox) or clicking a place on the map, the highest ranked news document with its title and the top 2 illustrative images is presented as a pop-up window positioning on the queried location of the map. The detailed ranking results about this location are shown in the right part with title, summary and

---

**Figure 5: The user interface of NewsMap.**

- New York mosque a ‘political football’
- Washington Post - Oct 20, 2010
- ABC News - Aug 18, 2010
- New York mosque a ‘political football’
- New York mosque a ‘political football’
- NewsMap
illustrative images (including the original pictures and the images enriched by our approach). In this way, the users can quickly get information about the news without reading its main body. Although UI is a very important part for such a system, a more detailed investigation is beyond our current scope and we leave it to our future work.

6. EXPERIMENTS

We elaborate our experiments on a dataset that contains news documents from multiple websites. We first introduce our experimental settings and then provide empirical justification of our approaches.

6.1 Experimental Settings

We collect a large data set from four websites, including ABCNews.com, BBC.co.uk, CNN.com and Google News. There are 135,308 news documents and 69,144 news images. By removing duplicates, there are 48,429 news documents and 20,862 news pictures in total.

From the news documents, there are 6,293 distinct location names extracted. By performing the filtering and expansion process, there are 4,742 locations remained.

To get the image similarity in Eq. 12, we extract 1000-dimensional bag-of-visual-words features from each image and adopt cosine similarity. In location relevance analysis, the rank $H$ of latent matrices is set to 100 and the parameters $\lambda_F$, $\lambda_L$, and $\lambda_J$ are set to 0.001, 0.001, 2$^{-3}$ and 2$^{-4}$, respectively. Notation that these values are small because the corresponding regularizers are not normalized. In image enrichment, we set $h = 20$, i.e., we collect the top 20 ranked results for each query. The parameters $c$ and $v$ are set to 5 (this number is set with considering multiple factors, such as the typical length of a news document and the space that the images occupy).

In empirical evaluation, we need to manually label several ground truths, including the relevance between a location and a news document and the relevance of an image with respect to a document. We establish three relevance levels: very relevant, relevant and irrelevant. For the relevance between location and news document, the manual labeling is mainly according to the following principles. “Very relevant” means that the location is an important information clue for the news, such as the news exactly happens at that place.

6.2 Experimental Results

6.2.1 On Location Relevance Analysis

We conduct two groups of experiments to evaluate our location relevance analysis.

Firstly, we randomly select 500 documents from the whole data set and then perform a location refinement process. For each document, we compare the relevance of locations before and after performing refinement. For the locations before performing refinement, we directly collect the distinct location names in the document. Supposing there are $l$ locations in the document, we also collect the $l$ locations with the highest relevance scores after performing CCPMF. Then each location of the news document is manually checked to decide whether it is really relevant. We assign score 2, 1, and 0 to indicate very relevant, relevant and irrelevant, respectively. In this way, we can obtain two average relevance scores for the locations in each document, one before and one after location refinement. Figure 6 illustrates the distribution comparison of the average scores, from which we can see that, for most news documents, the scores have been boosted after performing location refinement. The mean results of the scores before and after performing location refinement are 0.492 and 0.954, respectively. This indicates the effectiveness of our relevance analysis approach.

The second group of experiments validates the location relevance analysis by news search. We randomly select 100 location names from our toponym list. We then perform search in our news data set and rank the search results with

Figure 6: The comparison of the distribution of the average scores between before and after location refinement. We can see that for most news documents the locations after refinement are more relevant.

Figure 7: The search performance comparison of these three methods. “relevant” means that the news document is related to the news but its relationship with the major event of the news is not strong, and “irrelevant” means that, whether or not the location appears in the news document, it is not related to the major event of the news. For the relevance of image, the labeling principles are mainly as follows. “Very relevant” means that the image well describes the story of the event of the news. “Relevant” means that, although it may not match the event of the news, the image is able to provide information about several important elements of the news, such as the people and activity, and “irrelevant” means that the images do not provide information about understanding the news and its important elements.

We also involve several user studies. All the user studies are conducted with 30 users frequently reading news online. They come from two countries with age from 20 to 35.
the relevance scores obtained by CCPMF. We compare our approach with the following two methods:

1. BM25 ranking model \((b = 0.75 \text{ and } k_1 = 2.0)\) [17].
2. Probabilistic Matrix Factorization (PMF) model [27].

The comparison of our approach with the PMF algorithm is able to validate the effectiveness of the two correlation terms in Eq. 5.

To save manual labeling effort, we randomly select 1,000 documents with and without image enrichment. Each user is asked to freely browse news documents and compare these two versions. They can choose “better”, “much better” and “comparable” options for the comparison of these two versions. If the enriched images provide useful information, the users will regard the version after image enrichment better. Otherwise, if the enriched images are fairly irrelevant, users will judge that the original version is better as the images provide no information and they will be distinctive. We quantize the results as follows. We assign score 1 to the worse version and the other version is assigned a score 2, 3 and 1 if it is better, much better and comparable than this one, respectively. The average rating scores and the standard deviation values are illustrated in Table 1. From the results we can clearly see the preference of users towards the version after image enrichment.

We perform image enrichment on all the other news documents. Figure 9 provides several examples. We first evaluate our rank aggregation approach. We compare our approach with the following two methods:

1. Naive Search: Performing image search by regarding the whole document title as a query on Google image.
2. Naive Fusion: Performing image search with each term in the news title (instead of multi-term combinations) and then fusing the results as in 4.2.

To save manual labeling effort, we randomly select 1,000 news documents from the whole data set for evaluation. If for a query there is no result returned, we set its NDCG value to 0. Figure 10 illustrates the average NDCG comparison of these three methods. We can see that our approach remarkably outperforms the other two methods. This demonstrates the effectiveness of our query generation and image collection approach.

We also conduct a user study to compare the news documents with and without image enrichment. Each user is asked to freely browse news documents and compare these two versions. They can choose “better”, “much better” and “comparable” options for the comparison of these two versions. If the enriched images provide useful information, the users will regard the version after image enrichment better. Otherwise, if the enriched images are fairly irrelevant, users will judge that the original version is better as the images provide no information and they will be distinctive. We quantize the results as follows. We assign score 1 to the worse version and the other version is assigned a score 2, 3 and 1 if it is better, much better and comparable than this one, respectively. The average rating scores and the standard deviation values are illustrated in Table 1. From the results we can clearly see the preference of users towards the version after image enrichment. We also perform a two-way ANOVA test [20] and the results are illustrated in the table as well. The \(p\)-values demonstrate that the superiority of our version is statistically significant.

6.2.3 On the News Ranking in NewsMap

In the above experiments, we have evaluated the location relevance analysis and image enrichment components. Now we evaluate our ranking approach in the NewsMap system.
Table 1: The left part illustrates the average rating scores and standard deviation values from the user study on the comparison of news documents before and after image enrichment. The right part illustrates the ANOVA test results.

<table>
<thead>
<tr>
<th>After image enrichment</th>
<th>Before image enrichment</th>
<th>F-statistic</th>
<th>p-value</th>
<th>F-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.367 ± 0.669</td>
<td>1.100 ± 0.305</td>
<td>58.486</td>
<td>1.970 × 10^{-8}</td>
<td>0.313</td>
<td>0.999</td>
</tr>
</tbody>
</table>

Figure 11: The search relevance comparison of different ranking methods.

We compare our approach with the following methods:

(1) PRT: PageRank with content timeliness as the static ranking, i.e., the right part of Eq. 18 is set to date_{tk}.

(2) PRR: PageRank with location relevance as the static ranking, i.e., the right part of Eq. 18 is set to score_{tk}.

(3) BM25: ranking with the BM25 model [17].

We evaluate the relevance of searching results with the 100 queries as in Section 6.2.1. The comparison among PRT, PRR, BM25 and ours is shown in Fig. 11. It is observed that BM25 achieves the worst performance, and PRR achieves the best performance. Actually, our approach is very close to PRR, but our method and PRT are better in terms of timeliness, which is demonstrated in Fig 12. We make a simple statistic analysis about the number of documents happening in the latest week before the time we finished our data collection, and the average ratio over 100 queries for top d(d = 5, 10, 20, 50, 100) results is used to evaluate the news timeliness. From the both respects, i.e., relevance and timeliness, our method has more satisfied performance.

We then conduct user study to compare our approach with each of the three methods in terms of the interestingness of returned results, which is actually integrated the news relevance, timeliness and importance. In each pairwise comparison, the users are allowed to perform search with any location name and compare the two ranking lists. They can choose “better”, “much better” and “comparable” options for the comparison of each two ranking schemes based on the interestingness. We then quantify the results as follows. We assign score 1 to the worse ranking scheme and the other scheme is assigned a score 2, 3 and 1 if it is better, much better and comparable than this one, respectively. The average rating scores and the standard deviation values of the three comparisons are illustrated in Table 2 - 4, respectively. From the results we can see the preference of users towards our proposed ranking approach. We also perform a two-way ANOVA test [20], and the results are shown in the tables as well. It is observed that the performance of our method statistically significantly outperforms others.

Figure 12: The timeliness comparison of different ranking methods. The vertical axis is the average percentage of the latest news (in the latest week before our data collection process) of the top d results.

6.2.4 On the Comparison of NewsMap and Yahoo News Map

Now we conduct user study to evaluate the performance of the overall system and compare it with the Yahoo News Map system. Participants were required to give scores from 1 to 5 (greater score indicates better result) in terms of the following aspects while they freely searched and browsed the results: Convenience (the system should be convenient for users to search and browse news), Efficiency (it should cost users little time to learn the news content and comprehensive information,) and Usefulness (it is necessary for news retrieval system to provide useful information.). The results are illustrated in Fig. 13, which shows the mean scores and standard deviations. It can be observed that the proposed NewsMap system outperforms Yahoo News Map in different aspects.

7. CONCLUSIONS AND FUTURE WORK

This paper proposes methods that can estimate the relevance of locations for a given document and enrich it with web images. For location relevance analysis, we employ a novel matrix factorization method, and for image enrichment we propose methods to generate queries for search
8. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 60903146, 60833006 and 90920303), and 973 Program (Project No. 2010CB327905).

9. REFERENCES