IDIAP RESEARCH REPORT

FINDING GROUPS OF PEOPLE IN GOOGLE NEWS

Dhiraj Joshi [1] Daniel Gatica-Perez [2]

IDIAP–RR 05-68

NOVEMBER 2005

PUBLISHED IN
Proc. ACM Int. Conf. on Multimedia, Workshop on Human-Centered Multimedia (HCM), Santa Barbara, Oct. 2006.

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

1 Department of Computer Science and Engineering, Pennsylvania State University, USA
2 IDIAP Research Institute and Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland

IDIAP Research Report 05-68

FINDING GROUPS OF PEOPLE IN GOOGLE NEWS

Dhiraj Joshi Daniel Gatica-Perez

NOVEMBER 2005

PUBLISHED IN
Proc. ACM Int. Conf. on Multimedia, Workshop on Human-Centered Multimedia (HCM), Santa Barbara, Oct. 2006.

Abstract. In this paper, we study the problem of content-based social network discovery among people who frequently appear in world news. Google news is used as the source of data. We describe a probabilistic framework for associating people with groups. A low-dimensional topic-based representation is first obtained for news stories via probabilistic latent semantic analysis (PLSA). This is followed by construction of semantic groups by clustering such representations. Unlike many existing social network analysis approaches, which discover groups based only on binary relations (e.g. co-occurrence of people in a news article), our model clusters people using their topic distribution, which introduces contextual information in the group formation process (e.g. some people belong to several groups depending on the specific subject). The model has been used to study evolution of people with respect to topics over time. We also illustrate the advantages of our approach over a simple co-occurrence-based social network extraction method.

1 Introduction

Today technology plays a dual role of enhancing as well as understanding human life. Computers have revolutionized human interaction and redefined the meaning of staying in touch. New age communication entails exchange of large amounts of multimodal information generated by a host of devices at all times of the day. As a result, the popularity of gizmos like camera phones and the growth of the internet have created new challenges for data miners who wish to study and uncover human connectivity networks. It has been argued that camera phones have changed the way personal photography has been perceived, in the past [19]. Search engines like Google index billions of pages containing text, images, and video. Information as diverse as huge picture galleries to authoritative news articles are at the click of a mouse. Hyperlink structure of the Web represents yet another important network between people, places and organizations which search engines like Google use to assign authority to Webpages. Network characteristics of the WWW have been carefully studied [3]. A more human oriented work addresses the similarities and differences between sociability and usability in context of online communities [17]. Approaches to build a web of trust in social aspects have been proposed [7]. The need for socially aware computation has been emphasized [16].

Besides being a massive repository of information, the WWW hides another form of social network between people which link analysis may often fail to uncover. Consider the travel blogs of two photographers who visit similar places and have an interest in photography but who may never know about each other due to their disjoint virtual spaces. Such connections will inevitably be missed by social network discovery methods which rely on direct contact between people such as email exchange. In our work, we propose a computational approach to model world news content to reveal one such social network between key players in global news events.

The automatic discovery of groups of people and their relations is an area of research that, although studied in social network analysis for several years [21], has seen a sharp increase of interest given the existence- and in some cases, public availability- of large amounts of data exchanged via e-mails [9], and posted on professional websites, chat rooms, blogs, etc., from which social connectivity patterns can be extracted [12]. An important source of information about people and their connections is Web news. In particular, Google news[1] has become one of the richest access points to international news in terms of content and coverage (Fig. 1). Everyday this page displays representative text, pictures, and links to news stories deemed as most relevant by Google (Fig. 1). Links to stories from around 4500 international news sources can be obtained from here. Such an enormous coverage brings the advantages of providing complex multimedia content, and a better balance in terms of viewpoints (e.g. political and religious) across sources, which contrasts with existing datasets widely used in text analysis (e.g. the text-only Reuters-21578 collection composed of newswires from a single news source [10]).

The goal of this work is to discover and quantify the emerging social network among people who occur frequently in news, by quantifying similarities between them in the *context* in which they appear in news. Many existing approaches for group discovery rely on the assumption that people's connections are described by simple binary relations (i.e. a pair of people are either related or not) [14]. In contrast, we aim at discovering the group structure of people in news using not only their co-occurrence in the same document, but also the document content itself. We present a simple model that first discovers the topic structure of a news collection, and then finds groups of people according to the discovered topics. The use of language information to detect relations between people using probabilistic models is an emergent trend [11, 20] that has been largely motivated by the recent invention of probabilistic models for collections of discrete data [8, 5]. In a related work, an analysis of a social network emerging in news was recently reported in [15], with a different goal than ours, as it addressed the questions of whether the small-world and power-law degree-distribution properties - phenomena that have been recently observed in many complex networks [2] - appeared in social networks built from news articles. Importantly, the construction of the social network in [15] used binary co-occurrence information, not content as proposed here. In a different research line, multimedia news data has been used to learn correspondences between faces in images and names in photo captions [4]. In yet another work, multimedia sensors have been used to capture and study social interactions between people [6].

To the best of our knowledge, our work is one of the first studies on content-based discovery of social

[1] http://news.google.com

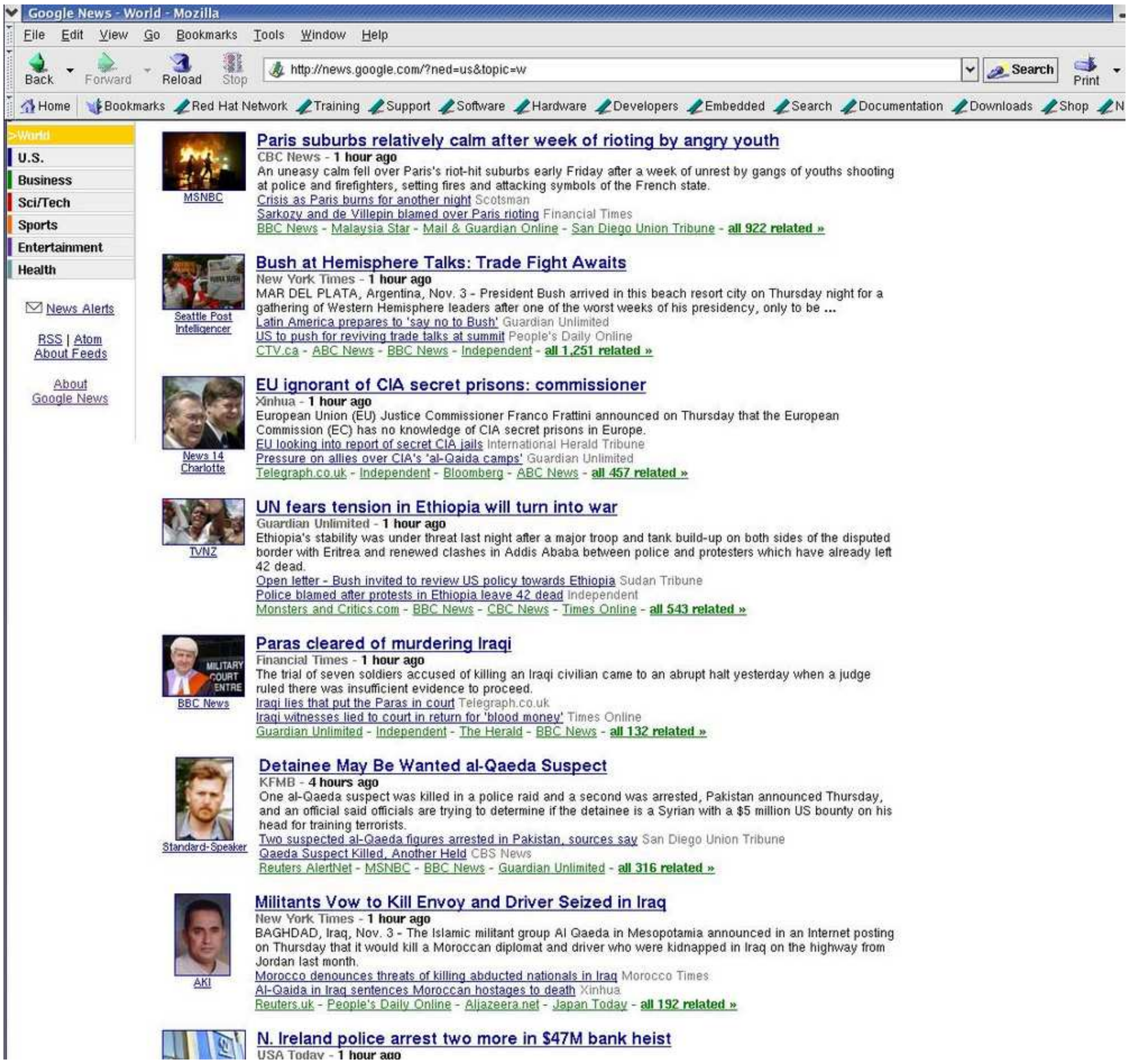

Figure 1: A snapshot of Google World News Page.

connectivity patterns using Web news data. Although our ultimate objective is to exploit social connectivity information existing in more than one media type (e.g. text and images), the approach presented here uses only textual information. Clearly, text is expected to constitute the strongest indicator of social relations in news, and the media type from which this information can be extracted more reliably. Later in the paper, we have discussed extension of the proposed approach to incorporate image information. A useful application of our work could be to discover soft links between travelers and photographers who maintain travel blogs containing text descriptions and pictures. Social network discovery using multimedia content can potentially reveal like minded people for future collaboration.

The remainder of the paper is organized in follows. In Section 2, we describe the data collection process and the nature of data collected. Section 3 describes our model to discover groups of people. In Section 4, we present the results and discussions. Section 5 provides some concluding remarks and discusses open issues.

(a) (b) (c) (d)

Figure 2: Four pictures obtained from Google news during the analyzed time period. The respective captions are (a) "Singh is escorted by Bush"; (b) "John Howard is standing by his criticism of a Melbourne cleric"; (c) "Islamic cleric Sheikh Mohammed Omran"; (d) "Complete coverage: London attacks."

2 News collection

2.1 News collection engine

The goal of our news collection engine is to begin from the Google world news page and collect news articles from links available on that page. Google automatically categorizes news stories, and the 20 most relevant stories are presented with some representative text, images and links to similar stories. Similar stories are potentially versions of the same event or closely related stories, as reported by other news agencies. Extracting hyperlinks from HTML pages involves some basic parsing operations. For simplicity, we implemented these procedures on our own. The news collection engine performed the following tasks:

1. From Google's world news page, text for each of the top 20 news daily stories were identified and stored.

2. Google provides links to the current available instances of each of the 20 news events. An instance corresponds to a version of the story or closely related news reported by a news agency. Our program obtained each available version. These were stored in raw HTML format. On average, around 1500 to 2000 news articles were collected and stored everyday by our crawler. Some of these could be copies of old news articles. The duplicates were later identified and removed, leaving a smaller set of documents for further processing.

2.2 News processing

A public-domain Java-based HTML parser [2] was used to extract text from HTML news files. News stories, collected over a period of about three weeks, formed our initial corpus for identifying an appropriate vocabulary to describe world news. Stopwords were identified and eliminated, stemming was performed, and words in news articles were ranked based on their frequency of occurrence to construct our vocabulary of 7413 words.

Additionally, we identified certain people occurring frequently in news. By observation, we discovered that proper nouns in text can be characterized by sequences of capital words (word beginning with an upper-case alphabet) followed by a lower case word. This characterization encompasses names of places, people, and news agencies. Therefore, a way of obtaining a list of people frequently occurring in news can be to track and obtain the frequencies of such sequences. Alternatively, one can also use a named-entity extraction program for the purpose. The obtained set was later manually trimmed to settle down upon a set of 32 people. The identified group of people, shown in Table 3, consists of politicians, terrorists, and heads of state of several nations, who appeared frequently in news between July 10, 2005 and July 31, 2005. As can be guessed from the list, the topics of interest during those weeks included the London bombings, the Israel-Palestine conflict, North Korea's nuclear program, Phillipines's political turmoil, etc. There are a few instances of the same person being spelt differently (e.g., *Kim Jongil* and *Kim Jong Il*), due to the lack of consistent naming among news agencies. In such cases, we treated differently spelt names, referring to the same person as different people.

[2] `http://htmlparser.sourceforge.net`

Person	Role
Mahmoud Abbas	Palestinian Authority President
Kofi Annan	UN Secretary General
Ian Blair	London Police Chief
Tony Blair	Britain Prime Minister
Bush	US President
Charles Clarke	Britain Home Secretary
Peter Clarke	London Police Anti-Terrorism Branch Head
Saeb Erekat	Palestine negotiator
Joseph Estrada	Phillipines deposed President
John Howard	Australia Prime Minister
Hasib Hussain	London Bomber
Saddam Hussein	Ousted Iraq President
Kim Jongil	North Korea Leader
Kim Jong Il	North Korea Leader
Laden	Al-Qaeda Leader
Gloria Macapagal Arroyo	Phillipines President
Ferdinand Marcos	Phillipines late President
Scott McClellan	US White House President Secretary
Shaul Mofaz	Israel Defense Minister
Abu Musab	Al-Qaeda Iraq Member
Pervez Musharraf	Pakistan President
Richard Reid	Shoe Bomber London
Condoleezza Rice	US Secretary of State
Ariel Sharon	Israel Prime Minister
Mohammad Sidique Khan	London Bomber
Mohammed Sidique Khan	London Bomber
Manmohan Singh	India Prime Minister
Jack Straw	Britain Foreign Secretary
Jalal Talabani	Iraq President
Shahzad Tanweer	London Bomber
Shehzad Tanweer	London Bomber
Nasser Yousef	Palestinian Authority Interior Minister

Figure 3: The table shows the 32 names (in alphabetical order) that were selected for study over a period of about six weeks, from July 10, 2005 to August 26, 2005. The corresponding roles of the people are also shown.

3 Finding topics and groups

The algorithm we present here consists of two stages. Each news story from the collection is considered as a document and represented by a bag-of-words [1]. In the first stage of the algorithm, a topic-based representation is automatically learned from the news collection. In the second stage, groups of people are automatically found using the topic-based representation for each of the documents in which a person's name appears. As outcome, the algorithm is able to assign probabilities to words (resp. people) as representing (resp. belonging to) different groups.

The first stage is implemented by applying Probabilistic Latent Semantic Analysis (PLSA) to the news corpus [8]. Due to lack of space, we briefly describe the PLSA process. Given a collection of D documents $\{d_i\}$ spanning a vocabulary of W words, PLSA models each word w_j in a document as arising from a mixture model. The mixture components are multinomial hidden variables z_k called aspects. A word can be generated by more than one aspect, and documents can thus be described by multiple aspects. Each w_j is conditionally independent of the document d_i it belongs to, given z_k. For K aspects, the term-document joint probability is given by

$$P(w_j, d_i) = P(d_i) \sum_{k=1}^{K} P(w_j \mid z_k) P(z_k \mid d_i). \tag{1}$$

With PLSA, a document d_i is thus characterized by a K-dimensional vector corresponding to its distribution over aspects $P(z|d_i)$, or in other words, by a low-dimensional topic-based representation. The model is learned in an unsupervised way via Expectation-Maximization (EM). Details about PLSA can be found in [8].

The second stage of the algorithm finds groups of people based on two basic assumptions: (1) people who belong to the same group can often be described as spanning the same topics (i.e., they are likely to have similar topic distributions), and (2) people often belong to more than one group (i.e, they can be described by more than one typical topic distribution). Assuming N people and M groups, let o_n and g_m denote the n-th person and m-th group, respectively. The algorithm uses the person-document co-occurrence information and the PLSA document representation as input, and outputs an estimation of the probability of each person of belonging to each of the groups, $P(g_m|o_n)$. A modification of K-means clustering is first applied on the document corpus using the Hellinger-Bhattacharya distance between PLSA document representations. Specifically, the distance between two documents d_i, d_{i^*}, represented by $P(z|d_i), P(z|d_{i^*})$, respectively, is computed as

$$\chi(d_i, d_{i^*}) = \left\{ \sum_{k=1}^{K} \left\{ \sqrt{P}(z_k|d_i) - \sqrt{P}(z_k|d_{i^*}) \right\}^2 \right\}^{1/2}. \tag{2}$$

After clustering the documents, the probabilities $P(g_m|o_n)$ are simply estimated as the fraction of documents in which person o_n appears that have been assigned to the cluster g_m,

$$P(g_m|o_n) = \frac{\sum_{i=1}^{D} 1[d_i \in g_m, d_i \circ o_n]}{\sum_{i=1}^{D} 1[d_i \circ o_n]}, \tag{3}$$

where $1[\cdot]$ is the indicator function, and $d_i \circ o_n$ denotes the binary person-in-document relation (i.e., person o_n appears in document d_i). Additionally, we are interested in estimating group-based word distributions, to be able to characterize each group by its most representative words (in probabilistic terms). We get an estimate of the word distribution given a group $P(w_j|g_m)$ by

$$P(w_j|g_m) = \sum_{k=1}^{K} P(w_j|z_k) P(z_k|g_m), \tag{4}$$

where the conditional distribution of an aspect given a group is computed by marginalizing over all people,

$$P(z_k|g_m) = \sum_{n=1}^{N} P(z_k|o_n, g_m) P(o_n|g_m), \tag{5}$$

and the conditional distribution $P(z_k|o_n, g_m)$ is computed by

$$P(z_k|o_n, g_m) = \sum_{i=1}^{D} P(d_i)P(z_k|d_i)1[d_i \circ o_n], \quad (6)$$

so the sum only considers the set of articles in which person o_n appears. As the outcome of the algorithm, each group g_m represents a set of documents having a specific mixture distribution over aspects. Clustering over the aspect distributions provides a formal representation for distinct news issues, some of which might not have been explicitly captured by individual aspects. In this sense, note that estimating word distributions per group allows to characterize a group by words that, although might potentially belong to quite distinct topics, are nevertheless representative of the group's topic mixture.

4 Experiments

4.1 Data and parameter setting

In experiments, news stories occurring in the six-week period described earlier, and containing at least one of the $N = 32$ people described in Table 1, were identified and converted to bags-of words, using our vocabulary of $W = 7413$ words. For learning the model from this data, documents which were too long or too short were removed, to reduce the effects of document size in the model. More specifically, we sorted each person's documents for each day, by their lengths. From this list, the first and third quartiles were determined, and only the news documents whose lenghts were in the inter-quartile range were kept for further processing. In the end, we were left with $D = 20799$ documents. Regarding the parameters of the model, unless stated otherwise, we arbitrarily set the number of aspects to $K = 7$, and the number of groups to $M = 5$. The choice of these parameters has obviously an impact in the performance of the model, but we did not explore ways of setting them automatically.

4.2 Group representation and membership

In order to characterize each group, we represent it by the top 10 words (ranked by $P(w|g_m)$), and the top 10 people (ranked by $P(o_n|g_m)$). Figures 4 and 5 show the five groups obtained (all figures show the stemmed version of words). By inspection, one can see that the top ranked words clearly identify the news issue which each group represents. The top ranked people, per group, are expected to be the key players with respect to such news topic. From common knowledge about world news, one can see that our model indeed seems to do so. Group 1 corresponds to *Palestine and Israel*, Group 2 corresponds to *Korea and nuclear issues*, Group 3 corresponds to *Iraq*, Group 4 corresponds to *London bomb*. The words associated with Group 5 are somewhat ambiguous. However, the associated people indicate that it could correspond to *Phillipines political turmoil*. It is also interesting, and somewhat intuitive, that a couple of people appear in the top rank of multiple groups. This could either be a result of people's active participation in multiple world issues given their political roles (e.g. George Bush and Condoleeza Rice, who appear top-ranked in groups 1-3), or the occurrence of indirect references to certain people in several world topics (e.g. Osama bin Laden, who appears top-ranked in Group 3, but also in Groups 1 and 5).

In a second experiment, with the intention of observing the discriminating ability of individual aspects, we performed document clustering using only a few aspects at a time, instead of the full aspect distribution. To do this, the Hellinger-Bhattacharya distance was calculated considering only one or two relevant aspects, and groups were constructed as described in Section 3. In this case, we fixed the number of groups to $M = 2$, to analyze if the algorithm could recover the group structure based on a "topic" vs. "non-topic" scenario. Figures 6 and 7 show the groups obtained when only the aspect probabilities corresponding to *London bomb* and *Israel-Palestine*, respectively, were used as features for clustering. In both cases, the probabilities $P(w|z)$ were used to identify the relevant aspects. For the case when only the *London bomb*-related aspects were used, the top 5 ranked people in the group related to this topic (as identified by the top ranked words) are indeed related to the subject, and also appeared in the group related to the London bombings obtained in our initial

Word	Prob.
gaza	0.0286
palestinian	0.0251
israel	0.0220
israeli	0.0218
said	0.0167
settlement	0.0122
settl	0.0103
bank	0.0093
west	0.0092
pullout	0.0076
Name	**Prob.**
Nasser Yousef	0.1813
Saeb Erekat	0.1810
Mahmoud Abbas	0.1727
Ariel Sharon	0.1721
Shaul Mofaz	0.1717
Condoleezza Rice	0.0611
Bush	0.0240
Scott McClellan	0.0114
Laden	0.0056
Kofi Annan	0.0046

Word	Prob.
nuclear	0.0236
said	0.0198
north	0.0195
iran	0.0187
korea	0.0169
talk	0.0163
unit	0.0080
stat	0.0071
korean	0.0070
south	0.0069
Name	**Prob.**
Kim Jong Il	0.2426
Kim Jongil	0.2422
Condoleezza Rice	0.0904
Scott McClellan	0.0874
Manmohan Singh	0.0805
Kofi Annan	0.0702
Bush	0.0617
John Howard	0.0413
Pervez Musharraf	0.0303
Jack Straw	0.0178

Word	Prob.
said	0.0235
iraq	0.0190
kill	0.0162
bomb	0.0115
attack	0.0114
iraqi	0.0111
sunni	0.0090
constitution	0.0082
baghdad	0.0074
police	0.0073
Name	**Prob.**
Jalal Talabani	0.2506
Abu Musab	0.2264
Saddam Hussein	0.1986
Scott McClellan	0.0568
Bush	0.0566
Laden	0.0505
Pervez Musharraf	0.0352
Condoleezza Rice	0.0243
Jack Straw	0.0195
Kofi Annan	0.0184

Figure 4: The figure shows groups 1, 2, and 3 (out of 5 groups), characterized by the words (ranked by $P(w|g_m)$) and people (ranked by $P(o|g_m)$).

Word	Prob.
bomb	0.0288
said	0.0273
london	0.0262
police	0.0246
attack	0.0160
suspect	0.0116
british	0.0090
arrest	0.0083
britain	0.0071
people	0.0070
Name	**Prob.**
Mohammed Sidique Khan	0.0911
Mohammad Sidique Khan	0.0910
Shahzad Tanweer	0.0907
Peter Clarke	0.0907
Hasib Hussain	0.0903
Richard Reid	0.0894
Shehzad Tanweer	0.0894
Ian Blair	0.0871
Charles Clarke	0.0641
Jack Straw	0.0499

Word	Prob.
said	0.0169
government	0.0082
minist	0.0065
lead	0.0065
people	0.0060
president	0.0058
country	0.0056
year	0.0054
world	0.0053
say	0.0049
Name	**Prob.**
Ferdinand Marcos	0.1281
Joseph Estrada	0.1279
Gloria Macapagal Arroyo	0.1198
John Howard	0.0856
Manmohan Singh	0.0835
Kofi Annan	0.0794
Tony Blair	0.0593
Laden	0.0447
Scott McClellan	0.0425
Pervez Musharraf	0.0402

Figure 5: The figure shows Groups 4 and 5 (out of 5 groups), characterized by the words (ranked by $P(w|g_m)$) and people (ranked by $P(o|g_m)$).

Word	Prob.
bomb	0.0273
said	0.0268
london	0.0246
police	0.0232
attack	0.0154
Name	**Prob.**
Richard Reid	0.0833
Mohammed Sidique Khan	0.0833
Mohammad Sidique Khan	0.0833
Shahzad Tanweer	0.0831
Shehzad Tanweer	0.0830

Word	Prob.
said	0.0184
gaza	0.0081
palestinian	0.0071
israel	0.0063
israeli	0.0062
Name	**Prob.**
Kim Jong Il	0.0498
Jalal Talabani	0.0498
Ferdinand Marcos	0.0498
Kim Jongil	0.0497
Ariel Sharon	0.0495

Figure 6: The figure shows the two groups obtained, characterized by the words (ranked by $P(w|g_m)$) and people (ranked by $P(o|g_m)$), using a constrained distance measure for clustering ,where only aspects related to *London bomb* were used in the distance.

Word	Prob.
gaza	0.0282
palestinian	0.0247
israel	0.0217
israeli	0.0215
said	0.0168
Name	**Prob.**
Saeb Erekat	0.1779
Nasser Yousef	0.1776
Ariel Sharon	0.1726
Shaul Mofaz	0.1722
Mahmoud Abbas	0.1711

Word	Prob.
said	0.0226
bomb	0.0147
police	0.0120
london	0.0116
attack	0.0092
Name	**Prob.**
Shehzad Tanweer	0.0379
Shahzad Tanweer	0.0379
Richard Reid	0.0379
Peter Clarke	0.0379
Mohammed Sidique Khan	0.0379

Figure 7: The figure shows the two groups obtained, characterized by the words (ranked by $P(w|g_m)$) and people (ranked by $P(o|g_m)$), using a constrained distance measure for clustering ,where only aspects related to *Israel-Palestine* were used in the distance.

Group 1	Group 2	Group 3	Group 4	Group 5
Abu Musab	Joseph Estrada	Charles Clarke	Ariel Sharon	Ferdinand Marcos
Bush		Hasib Hussain	Condoleezza Rice	Gloria Macapagal Arroyo
Jalal Talabani		Ian Blair	Mahmoud Abbas	
John Howard		Jack Straw	Nasser Yousef	
Kim Jongil		Mohammad Sidique Khan	Saeb Erekat	
Kim Jong Il		Mohammed Sidique Khan	Shaul Mofaz	
Kofi Annan		Pervez Musharraf		
Laden		Peter Clarke		
Manmohan Singh		Richard Reid		
Saddam Hussein		Shahzad Tanweer		
Scott McClellan		Shehzad Tanweer		
		Tony Blair		

Table 1: Groups of people obtained using a spectral clustering method and co-occurrence-only data.

experiment (group 4 in Figure 4). Note also that there were some differences in specific rankings. The other group (shown also in Fig. 6) is quite mixed in topics and people. It is interesting to note that none of the people in the first group appeared as top-ranked in the second one. A similar trend can be observed for the case when only the *Palestine-Israel*-related aspects were used. Keeping in mind that that the number of aspects in PLSA in our experiments is rather small -so roughly speaking, each individual aspect relates to a specific news topic-, the result of these experiments highlight a fact: many news stories are mainly about a single topic, and thus people that make those news also relate to mainly one topic. An example of these situation could be the London terrorists. However, as our first experiment suggested, there are other people who naturally belong to different groups given the multiple events they are involved in.

Our method has the advantages of using content, rather than only binary co-occurrence information, to find groups, and of being able to assign probabilities of group membership to people. As an initial way of comparing our approach, in a third experiment we clustered people into groups using only co-occurrence data, i.e. names appearing on the same news article. Let $\mathcal{N}(n)$ and $\mathcal{N}(n^*)$ denote the number of news documents in which people o_n and o_{n^*} occur, respectively, and $\mathcal{N}(n, n^*)$ denote the number of documents in which the two people cooccur. A pair-wise similarity measure between people o_n and o_{n^*}, $s(o_n, o_{n^*})$ was then defined as

$$s(o_n, o_{n^*}) = \frac{\mathcal{N}(n, n^*)}{min(\mathcal{N}(n), \mathcal{N}(n^*))}.$$

This similarity measure was then used as input to a spectral clustering algorithm to group people [13]. The algorithm constructs a pair-wise people similarity matrix. After matrix pre-processing, its spectrum (eigenvectors) is computed, the M largest eigenvectors are stacked in columns in a new matrix, and the rows of this new matrix are normalized. Each row of this matrix constitutes a feature associated to each person. The rows of such matrix are then clustered using K-means (with M clusters), and all people are labeled accordingly, which produces a hard partition of the list of people. Details of the spectral clustering algorithm can be found in [13]. In this experiment, we clustered people into $M = 5$ groups, as before. The groups obtained are shown in Table 1. We can observe that, although some of the clusters are quite meaningful, e.g. group 3 relates to *London bomb*, group 4 to *Israel-Palestine*, and group 5 to *Phillipines*, the other groups turned out to be either very mixed or too small. Furthermore, some people that could naturally belong to several groups are assigned to only one of them by the spectral method. This is an inherent limitation of any hard-clustering approach.

4.3 Studying people's topic evolution

As an application of our framework, we looked at topic evolution with respect to a few multi-role key players in news. Specifically, for a given person, each word in their news articles was assigned to a group, based on the distribution $P(g_m|w)$. The temporal scale was divided into disjoint windows of five days, and the fraction of words corresponding to each person and assigned to each group was calculated over each time period. This fraction can be thought to represent the extent to which a particular person was involved in a certain news issue [18]. Figure 8 shows the results for six people. The legends in the graphs indicate the groups to which trends correspond to. We plot the trend for the four most prominent groups, namely *Palestine, Iraq, London bomb, and North Korea*, and identify them by the respective terms.

We now briefly discuss interesting trends with respect to individual people. In order to explain the trend, we refer to the fraction of words assigned to a certain group for a given person as the respective topic itself.

- **George W. Bush** - The topic *London bomb* is high and drops over time, which is expected, as it is a relatively short-lived issue. Topics *Korea* and *Iraq* fluctuate over time. Interestingly, we notice a high value of topic *Palestine* around the end, which is expected to be the approximate period of the *Gaza pullout* event.
- **Condoleeza Rice** - There is a relatively low value of the topic *London bomb*, all the time. However, we notice a peak in *Palestine* at the same time as for *George W. Bush* (approximate time of *Gaza pullout*). Topic *Korea* is particularly high, especially around the beginning and the end.
- **Tony Blair** - We notice a very high value of *London bomb*, fluctuating over time. This is an expected result. A peak in *Palestine* is noticed around the end, consistent with Condoleeza Rice and George Bush.
- **Saddam Hussein** - The topic *Iraq* dominates throughout which is again an expected result. The remaining topics do have small values throughout. This could be primarily because of the inherent similarity in the topics, such as concerns about security, nuclear threats, terrorism, and hence a somewhat similar use of language.
- **Pervez Musharraf** - The topic *London bomb* is high and drops over time. Topic *Iraq* fluctuates over time. A small peak in *Palestine* is noticed towards the end. A particularly high value of *Korea* is seen. Although Mr. Musharraf has no direct relation to the subject, this peak could again be explained as the result of similar use of language and similar concerns across the topics.
- **Osama bin Laden** - The topic *London bomb* is high and drops over time. A peak in *Palestine* is noticed towards the end. The rest of the topics fluctuate over time.

4.4 Classification of new documents

For each group, a representation was obtained as the centroid of all documents' aspect distributions $P(z|d)$ which form the particular group, and computed during training. A classifier for new documents containing at least one of the people listed in Table 3 was set up as follows. Each group was first identified by the $\mathcal{N}$ top ranked people, where people were ranked based on the probabilities $P(o_n|g_m)$. Then, for a new document d_{new}, a group $g_{m^{new}}$ was first assigned using nearest neighbor criterion. The distance used here was the Hellinger-Bhattacharya distance, as explained earlier. During classification, each occurrence of a person was treated as an instance of the document. For the new document d_{new}, classification to group $g_{m^{new}}$ was deemed correct if the corresponding person $o_{n^{new}}$ was one of the $\mathcal{N}$ representatives of $g_{m^{new}}$. The classification performance for different values of $\mathcal{N}$ is shown in Figure 9. The total number of documents used for testing purposes was 3157 (the set is obviously disjoint from the documents used for training). We observe that this simple approach can correctly predict documents as containing the correct person -based on the documents' content- in more than 80% of the cases when $\mathcal{N} \geq 10$. This is an interesting result given that the method is fully unsupervised, but would of course need to be validated against a standard supervised approach in order to assess its comparative performance.

5 Discussion and Conclusion

We studied the problem of social network discovery from Web news data. Groups were constructed based on the representation of news stories as mixtures of aspects -using probabilistic latent semantic analysis-, and a simple probabilistic framework to associate people with groups was proposed. This was used to quantify the involvement of people in different groups over time. The study of both the discovered groups and the evolution of topics was coherent with our common knowledge about world events, but is however more difficult to assess objectively. A systematic objective evaluation of our approach is clearly a non-trivial issue that requires further

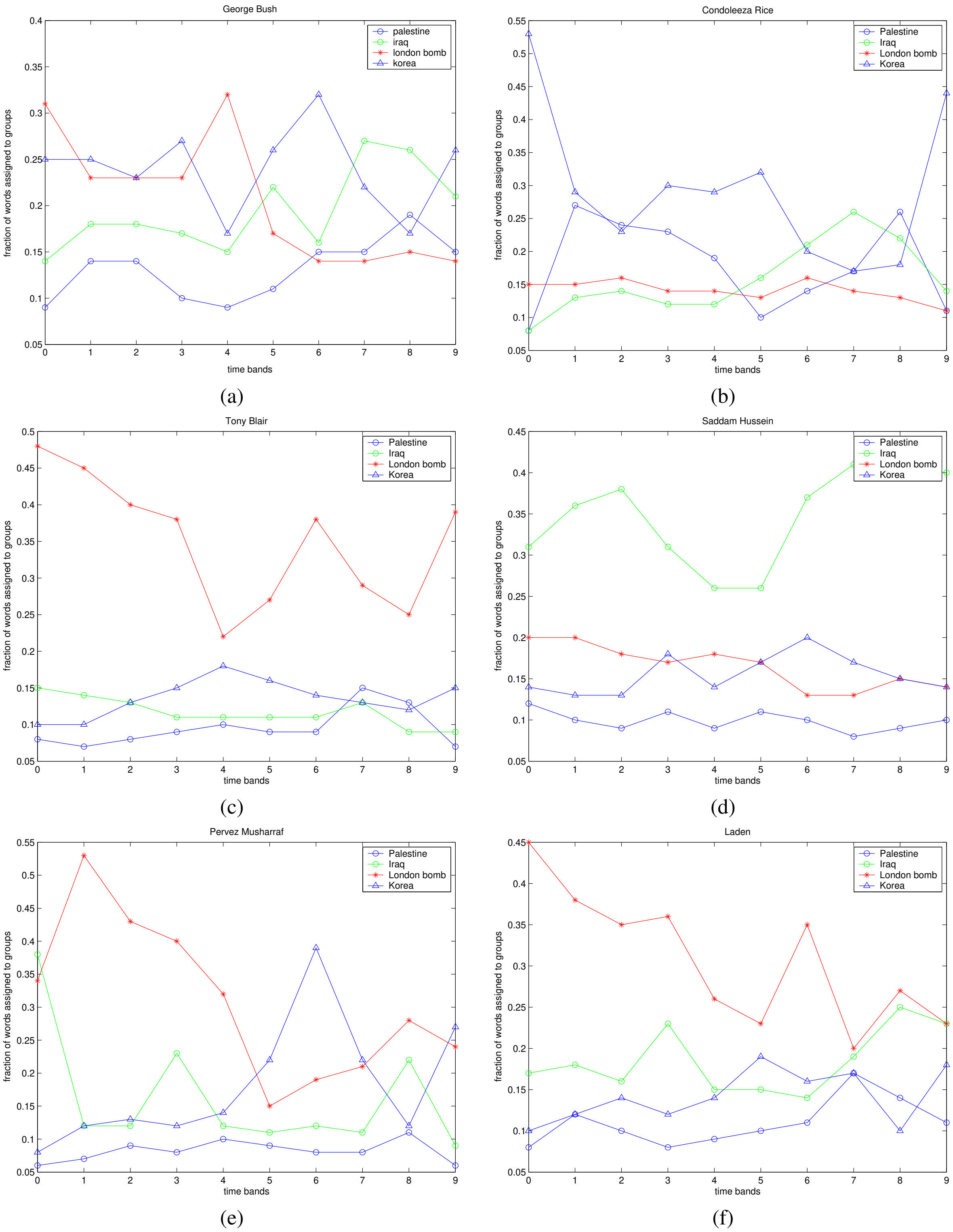

Figure 8: The figure shows topic evolution with respect to certain people, over a period of about six weeks. In each graph, the X axis represents the time period, and the Y axis represents fraction of words per group. The people are (a) George Bush, (b) Condoleeza Rice, (c) Tony Blair, (d) Saddam Hussain, (e) Pervez Musharraf, and (f) Osama bin Laden.

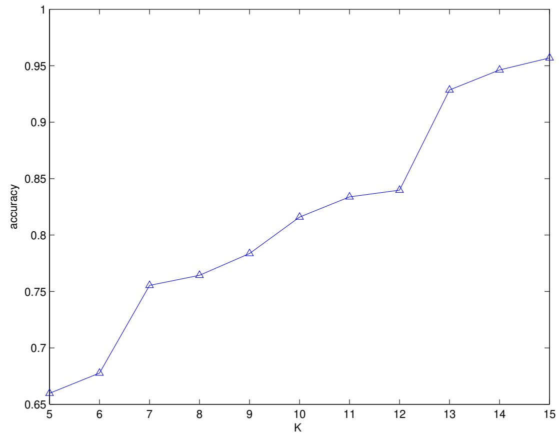

Figure 9: Classification performance of the algorithm for different values of parameter $\mathcal{N}$, i.e., the number of representatives each group g_m is represented by.

work. Such an evaluation would include the effect of varying the parameters that were kept as part of this study. The problem of scalability to larger number of people and amount of news also needs to be studied.

One potential extension of this work could be to incorporate image information, leading to the construction of a social network from multimedia data. We could consider the images embedded in Web news articles and design a system to learn a social network between commonly occurring people from image data. Automatic face detection/recognition methods would be necessary for this purpose. This model could then be integrated with our current work, which uses only text. It would indeed be interesting to study the coherence of social networks learned using text and image data.

However, there are a few challenges involved in this direction. In the first place, automatic image collector programs which search the Web often bring a very large number of unrelated images, e.g. logos, icons, and advertisements. In order to perform analysis, an effective (and maybe semiautomatic) algorithm for pruning this dataset would likely be required. Issues to evaluate include whether the number of relevant images collected in this fashion would be sufficient to learn a meaningful social network, and whether visual processing algorithms would be robust enough for reliable person identification. In the second place, relevant images occurring within news stories often contain no people at all. Finally, certain people occurring frequently in news text could be seldom depicted in pictures. All these issues have to be considered for future work in this domain.

6 Acknowledgement

This work was supported by the Swiss NCCR on Interactive Multimodal Information Management (IM2). The authors would like to thank Florent Monay and Dong Zhang for insightful discussions during this work.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval," ACM Press, 1999.

[2] A. L. Barabasi, "Linked: The New Science of Networks," Perseus, 2002.

[3] A. L. Barabasi, R. Albert, H. Jeong, "Scale-free Characteristics of Random Networks: the Topology of the World Wide Web," Physica, 281, 69–77, 2000.

[4] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E.L. Miller, and D. Forsyth, "Faces and Names in the News," Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR), 2004.

[5] D. Blei, A. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, 993–1022, 2003.

[6] T. Choudhury and A. Pentland, "Sensing and Modeling Human Networks using the Sociometer," Proc. Int. Conf. on Wearable Computing, 2003.

[7] J. Golbeck, B. Parsia, and J. Hendler, "Trust Networks on the Semantic Web," Lecture Notes in Computer Science, 238–249, 2003.

[8] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," Machine Learning, vol. 42, 177–196, 2001.

[9] B. Klimt and Y. Tang, "The Enron Corpus: A New Dataset for Email Classification Research" Proc. European Conf. on Machine Learning, 2004.

[10] D. Lewis, "Reuters-21578 Test Categorization Test Collection," `http://kdd.ics.uci.edu/databases/reuters21578.html`, 1997.

[11] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "Topic and Role Discovery in Social Networks," Proc. Int. Joint Conf. on Artificial Intelligence, 2005.

[12] M. E. J. Newman, "Coauthorship Networks and Patterns of Scientific Collaboration," Proc. National Academy of Sciences, 2004.

[13] A. Ng, M. I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," Proc. Neural Information Processing, 2001.

[14] K. Nowicki and T. A. B. Snijders, "Estimation and Prediction for Stochastic Blockstructures," Journal of American Statistical Association, vol. 96, 1077–1087, 2001.

[15] A. Ozgur and H. Bingol, "Social Networks of Co-occurence in News Articles," Lecture Notes in Computer Science, 3280, 688–695, 2004.

[16] A. Pentland, "Socially Aware Computation and Communication," IEEE Computer, vol. 38, 33–40, 2005.

[17] J. Preece, "Online Communities: Designing Usability and Supporting Sociability," John Wiley & Sons, 2000.

[18] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The Author Topic Model for Authors and Documents," Proc. 20th conference on Uncertainty in Artificial Intelligence, 2004.

[19] N. A. Van House and M. Davis, "The Social Life of Cameraphone Images,". Proc. Pervasive Image Capture and Sharing Workshop at the Seventh Int. Conf. on Ubiquitous Computing, 2005.

[20] X. Wang, N. Mohanty, and A. McCallum, "Group and Topic Discovery from Relations and Text," Proc. KDD Workshop on Link Discovery: Issues, Approaches and Applications, 2005.

[21] S. Wasserman and K. Faust, "Social Network Analysis: Methods and Applications," Cambridge University Press, 1994.