Relationship Discovery in Large Text Collections Using Latent Semantic Indexing

R. B. Bradford
SAIC, Reston, Virginia
bradfordr@saic.com

Abstract

This paper addresses the problem of information discovery in large collections of text. For users, one of the key problems in working with such collections is determining where to focus their attention. In selecting documents for examination, users must be able to formulate reasonably precise queries. Queries that are too broad will greatly reduce the efficiency of information discovery efforts by overwhelming the users with peripheral information. In order to formulate efficient queries, a mechanism is needed to automatically alert users regarding potentially interesting information contained within the collection. This paper presents the results of an experiment designed to test one approach to generation of such alerts. The technique of latent semantic indexing (LSI) is used to identify relationships among entities of interest. Entity extraction software is used to pre-process the text of the collection so that the LSI space contains representation vectors for named entities in addition to those for individual terms. In the LSI space, the cosine of the angle between the representation vectors for two entities captures important information regarding the degree of association of those two entities. For appropriate choices of entities, determining the entity pairs with the highest mutual cosine values yields valuable information regarding the contents of the text collection. The test database used for the experiment consists of 150,000 news articles. The proposed approach for alert generation is tested using a counterterrorism analysis example. The approach is shown to have significant potential for aiding users in rapidly focusing on information of potential importance in large text collections. The approach also has value in identifying possible use of aliases.

Key Words: information discovery, latent semantic indexing, LSI, entity extraction, counterterrorism analysis

1 Introduction

There is an increasing need for methods for efficiently identifying information of interest in large collections of text.[*] For example:

- Corporations wish to survey large collections of e-mails to identify illicit activities or those that may potentially generate liability.
- Legal firms need to review large collections of documents obtained during the discovery phase of litigation.
- Counterterrorism analysts have a requirement to rapidly review large amounts of incoming messages to identify information relevant to potential threats.

In general, humans must read at least some of the documents contained in collections in order to make definitive judgments about the presence or absence of new and interesting information. For large text collections, the key problem is to minimize the amount of information that must be read by a user in order to make such judgments. The key problem is that the user typically has only a very general idea of the specific contents of the collection. In order to efficiently select documents for analysis, the user must invoke reasonably precise queries. Queries that are too broad will yield large amounts of material that may be only peripherally relevant. What is needed is a method for users to focus their attention on specific topics within the collection that are of interest to them.

Generally speaking, users are interested in relationships among entities. Entities of particular interest typically include people, organizations, and locations. Relationships of interest may include communications, financial

[*] For text databases, large is in the eye of the beholder. Generally speaking, however, as text collections grow beyond 100,000 documents, classical text analysis techniques become increasingly cumbersome.

transactions, and physical proximity. Items of interest may consist of complex combinations of entities and relations, such as events. Users typically have some entities and relationships that they know are of interest. However, this information is almost always incomplete. Any method for providing tip-off of items of potential interest must allow for the appearance of new and unexpected entities and relationships.

Latent semantic indexing is a well-established method for extracting relationship information from large collections of text [3]. In most LSI applications, the emphasis has been on comparison of documents. However, the LSI technique generates representation vectors not only for the documents of a collection, but also for the elements of which they are comprised. In most LSI applications, the constituent elements used in creating the LSI space are the individual terms that occur in the documents of the collection.[†] In the experiment described here, the constituent elements of the documents that are incorporated into the LSI indexing consist of a mixture of individual terms and multi-term sequences corresponding to named entities. Treating named entities as units in creating the LSI representation space yields important advantages. In particular, the representation space that is generated codifies relationships among the entities per se. The experiment described here was designed to test the utility of such representation in identifying topics of possible interest within large text collections.

For any two entities present in the text collection of interest, the proximity of their representation vectors in the LSI space provides a direct measure of their degree of contextual association in the collection. This feature can be exploited to provide rapid overviews of the aggregate implications of the relations present in large collections of unstructured information. In this work, two-dimensional cross-correlation matrices are used to highlight potentially interesting associations between entities. Identification of such associations allows users to focus their attention on information in the collection that may be of particular interest. This is of great utility in facilitating rapid overview of large quantities of text.

2 Methodology

Previous work has demonstrated the utility of applying matrix decomposition techniques for extraction of relationship information from text collections. In particular, a comprehensive series of papers by Skillicorn has demonstrated the potential of this approach [9]-[12]. Skillicorn notes four aspects of matrix decompositions that make them particularly attractive for analysis of terrorist networks:

- They incorporate higher-order correlation information.
- They can exploit auxiliary information associated with both edges and nodes of social networks.
- They are robust in the presence of missing and incorrect data.
- They scale much better than many tools classically used in social network analysis and link analysis.

In the experiment described here, entity extraction and the matrix decomposition technique of latent semantic indexing were combined to allow direct analysis of entity-entity relationships.

This investigation employed a database of 158,492 English-language news articles from the time period 2002 through 2003. As hardcopy, this would constitute approximately 60 boxes of documents. The collection contained 332,386 unique terms.

The text of the articles was pre-processed using the LingPipe entity extraction software.[‡] All named entities extracted by the LingPipe software were treated as indexing units in the creation of the LSI representation space. This was done using the phrase processing option of the LSI indexing software employed.[§] In the creation of the LSI space, log (1+term frequency) was used for local weighting and entropy for global weighting, with the number of dimensions set to 300. A stopword list of 571 words was employed.

The use case chosen for the experiment was that of an intelligence analyst needing to rapidly overview a collection of text to identify information related to potential threats.

[†] There have been studies of the use of phrases as indexing elements in creating LSI spaces. What is different here is a specific focus on the use of named entities as indexing elements.

[‡] LingPipe is a suite of Java tools designed to perform linguistic analysis on natural language data. The software can be downloaded from: http://www.alias-i.com/lingpipe/

[§] The LSI space was created using version 2.6.2 of the text analytics software from Content Analyst Company.

Threat scenarios can involve a wide variety of entities and relationships. The most interesting entities in such scenarios are individual terrorists, terrorist groups, targets, and weapons. Thus, the analysis focused on those entities.

Using available training data, LingPipe was not capable of directly identifying these types of entities per se. For the purposes of this experiment, terrorist groups were identified as such by cross-correlating the list of entities tagged by LingPipe as organizations with a list of known terrorist groups. Some (potential) targets were identified as such through cross-correlation of a list of potential targets (the Pentagon, Eiffel Tower, etc.) with the list of entities tagged by LingPipe as locations. Other targets were identified within the list of location entities generated by LingPipe based on the appearance of terms from a list of descriptors associated with potential targets (e.g., airport, refinery, church, etc). Weapons were identified based on cross-correlation with a list of chemical and biological weapons.

Both documents and indexing units are represented by vectors in the LSI space. The LSI software can provide the cosine of the angle between any two such representation vectors. As noted above, the indexing units in this case consisted of both individual terms and sequences of individual terms recognized as meaningful units by the entity extraction software. Cross-correlation matrices were generated for a number of combinations of people, groups, targets, and weapons. Each entry in these matrices is the cosine of the angle between the representation vectors in the LSI space for the corresponding row and column entities. In the LSI space, higher cosine values are indicative of stronger relationships. For each matrix, the highest cosine entries were investigated.

In the experiment, a number of the identified associations of interest were related to the Salafist Group for Call and Combat (GSPC). This group originated in Algeria, but has developed an extensive network in Europe. It has close ties to Al Qaeda. The group was quite active in Europe during the time frame covered by the documents in the test collection.

3 Latent Semantic Indexing

The key element of the experiment reported here is the application of the technique of latent semantic indexing. This technique was used to automatically determine the relationships among terms and, more importantly, named entities in the text collection.

The technique of latent semantic indexing consists of the following primary steps [3]:

1. A matrix is formed, wherein each row corresponds to a term that appears in the documents of interest, and each column corresponds to a document. Each element (*m*,*n*) in the matrix corresponds to the number of times that the term *m* occurs in document *n*.
2. Local and global term weighting is applied to the entries in the term-document matrix.
3. Singular value composition (SVD) is used to reduce this matrix to a product of three matrices, one of which has non-zero values (the singular values) only on the diagonal.
4. Dimensionality is reduced by deleting all but the *k* largest values on this diagonal, together with the corresponding columns in the other two matrices. This truncation process is used to generate a *k*-dimensional vector space. Both terms and documents are represented by *k*-dimensional vectors in this vector space.
5. The relatedness of any two objects represented in the space is reflected by the proximity of their representation vectors, generally using a cosine measure.

LSI has some particularly attractive features for information discovery:

- LSI vectors reflect subtle relationships derived from a holistic analysis of the totality of information indexed [6].
- There is complete generality of the approach with regard to subject matter, genre, and language. In fact, documents in multiple languages can be analyzed simultaneously [7].
- Identified relationships represent the aggregate implications of high-order associations. For this reason, the approach is capable of identifying quite subtle relationships. Routinely, strong (and correct) relationships are shown to exist where there are, in fact, no first-order associations in the documents that are indexed [5].
- In at least some applications, the effectiveness of the technique increases as the amount of data increases [1].

4 Results

4.1 Entity-entity comparisons – terrorist groups versus targets

Table 1 is an LSI-derived matrix presentation of relationship information between terrorist groups and potential targets for this collection of text. The groups and targets shown are a sampling of such entities that appeared among the 158,492 news articles and that were identified as entities by the LingPipe software. Each entry in the matrix is the cosine of the angle between the representation vectors for the respective group and target.

The great majority of the entries in Table 1 show a very low degree of relatedness, with typical cosine values in the range of -.1 to +.1. This is typical of objects in an LSI representation space that have a very low degree of association. Dennis et al have reported LSI cosine values for random pairs of words as .02 ± .03 [2]. Figure 1 shows the distribution of cosine values between term representation vectors for the type of material treated here.

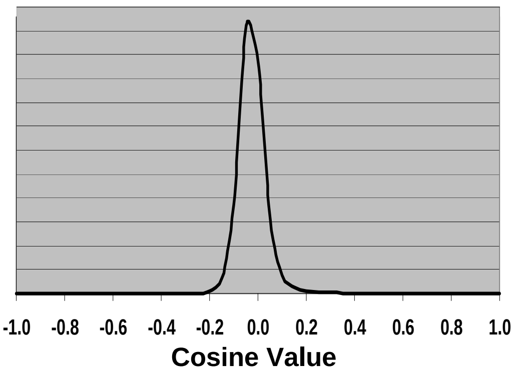

Figure 1: Distribution of cosine values between LSI term representation vectors

The cosine value of .5087 between the vector corresponding to the *GSPC* and that for *Strasbourg Cathedral* is far larger than the other entries. This is a correct association. The Frankfurt cell of the GSPC indeed had planned an attack on the cathedral in Strasbourg to take place at Christmas of 2000. In this document collection, *no* articles directly link the term GSPC (or any synonym for GSPC) with Strasbourg Cathedral. This is a good example of the fact that the LSI technique takes into account higher-order relations in making associations. In this collection, two articles associate the names of specific individuals with Strasbourg Cathedral. Relationships among names in other articles, in aggregate, associate these individuals

Table 1: Group vs. target matrix

	Athens Airport	Gdansk Oil Refinery	NATO Headquarters	The Pentagon	Strasbourg Cathedral	Prague Castle	Trafalgar Square	The Vatican
Abu Sayyaf Group	-.0254	.0236	-.0235	.0292	-.1235	.0231	-.0862	.0805
Al Aqsa Martyrs Brigade	-.0152	-.0117	-.0572	-.0343	-.0710	.0072	-.0120	-.0222
Al-Takfir Wa Al-Hijrah	-.0562	.0576	-.0177	-.0243	.0919	-.0015	-.0346	.0009
Ansar al-Islam Group	-.0086	.0074	-.0048	-.0911	.0925	.0044	-.0733	-.0826
GIA	-.0349	-.0055	-.0122	-.0291	.3646	.0156	.0598	-.0287
GSPC	-.1556	.0499	-.0043	-.1045	**.5087**	.0261	.1677	-.1343
HAMAS	.0071	-.0262	.0462	.0521	-.0019	.0271	.0152	.0065
Hizballah	-.0513	.0349	.0028	-.0595	-.0508	-.0169	-.1056	.0458
Islamic Jihad	.0138	-.0409	.0300	.0281	-.0439	.0253	.0086	-.0133
Jamaah Islamiyah	-.0458	.0618	-.0026	.0053	.1530	.0001	.0013	-.0408
Mojahedin-e Khalq Organization	-.0033	.0554	.0158	.1217	.0112	-.0016	-.0578	-.0616
Palestine Liberation Front	.0195	.0329	-.0220	.0147	-.0088	-.0128	.0685	.0308
PFLP	.0580	-.0107	-.0190	.0135	.0101	-.0115	.0931	.0365
Salafia Jihadia	.0101	.0317	-.1304	-.1537	.2928	-.0102	-.0264	.0393

with the GSPC. The high cosine value here thus is based completely on nth-order associations. Derivation of relationships based entirely on higher-order associations is a common occurrence in LSI processing. Kontostathis has shown that LSI can derive relationship information from such associations up to at least fifth order [6]. Although the individual contributions of higher-order terms may be relatively small, this is compensated for by the fact that there are so many of them. In a collection of six thousand documents, Kontostathis found fifty thousand first-order co-occurrences of noun phrases, but over ten million second-order co-occurrences and over sixty million third-order co-occurrences [6].

4.2 Entity-entity comparisons – terrorist groups versus weapons

Table 2 presents a matrix of terrorist groups versus weapon type, produced in the same manner as Table 1. As in Table 1, the majority of cosine entries in this matrix are very small.

One entry stands out among the rest - the cosine value of .3375 corresponding to the entities GSPC and ricin. This is a correct association. GSPC personnel planned attacks using ricin during the time frame covered by the documents of the test collection. Once again, there are no documents in the collection that contain both the term GSPC and the term ricin. However, there are articles that contain the term ricin together with the names of several individuals. These names, in turn, appear in articles that contain the term GSPC (or one of its synonyms).

The next largest cosine value in the table is .3314 between the representation vectors for Jamaah Islamiyah and cyanide. This also is a correct association. In this time frame traces of lead cyanide as well as a manual describing the use of cyanide in attacks were found at a raided Jamaah Islamiyah hideout.

The third largest cosine in the table is .2825, between the representation vectors for GSPC and cyanide. This is another correct association. In the time frame covered, the GSPC planned attacks using cyanide.

Table 2: Group vs. weapon matrix

	Anthrax	Botulinum	Cyanide	Phosgene	Ricin	Sarin	Smallpox	Tularemia
Abu Sayyaf Group	-.0078	.0388	.1300	.0045	-.1500	.0220	-.0382	-.0014
Al Aqsa Martyrs Brigade	-.0038	.0003	.0601	.0301	-.1031	.0578	-.0611	.0085
Al-Takfir Wa Al-Hijrah	-.1159	-.0217	.1372	.0077	.0096	-.0485	-.1303	-.0053
Ansar al-Islam Group	.0430	.0741	.1220	.1368	-.0059	.0913	-.0381	-.0128
GIA	.0250	-.0572	.1055	-.0536	.1722	-.0573	-.0466	.0250
GSPC	-.0168	.0159	.2825	-.0129	**.3375**	-.0237	.0465	.0116
HAMAS	.0301	.0636	-.0880	.0094	.0023	.0117	.0294	.0101
Hizballah	-.0993	-.0137	-.0470	-.0338	-.0563	-.0138	-.0673	-.0143
Islamic Jihad	-.0032	.0290	-.0048	-.0524	-.0747	-.0109	-.0041	.0091
Jamaah Islamiyah	-.0344	.0807	.3314	.0464	.1002	-.0064	-.0528	-.0040
Muhajedin-e Khalq Organization	.0124	.0078	-.0636	.0016	-.0722	.0139	.0287	-.0137
Palestine Liberation Front	.0128	.0046	.0214	.0354	.0247	-.0125	-.0440	.0142
PFLP	.0748	.0891	.0730	.0401	-.0316	.0162	.0264	-.0044
Salafia Jihadia	-.1427	-.0875	.0648	-.0552	.1059	-.1080	-.1108	.0001

4.3 Entity-entity comparisons – terrorist groups versus names

Relationships between people and terrorist groups are of particular interest. A matrix presentation of group versus person yields direct information regarding such relationships. For this collection, that matrix is quite large. Table 3 shows data extracted from a row in this matrix. The selected row corresponds to the GSPC. The names in this row with the ten highest cosine values are shown in the table.

Table 3: Ten most closely associated names for the named entity GSPC

	Name	Cosine
1	Hassan Hattab	.9632
2	Abou Hamza	.8852
3	Antar Zouabri	.8641
4	Mohammed Slim	.8200
5	Kamel Daoudi	.8198
6	El Bassir	.8185
7	Djamel Beghal	.8078
8	Abdelrazzaq el Mizalli	.7994
9	Abassi Madani	.7958
10	El Para	.7604

All ten of the cosines in Table 3 are large. This implies that a close association exists between these individuals and the GSPC. That is a correct implication. Hassan Hattab was the founder of the GSPC. Abou Hamza is the primary alias used by Hassan Hattab. This is an example of the coalescence of true names and aliases when using LSI indexing. Antar Zouabri was the leader of the GIA, the Algerian terrorist group from which the GSPC broke off. Mohammed Slim is an Algerian terrorist with very close connections to the GSPC. Kamel Daoudi is a key member of the GSPC Beghal cell. El Bassir is the alias used by Mohamed Lounis, chief of external relations for the GSPC. Djamel Beghal is the leader of a cell of the GSPC. Abdelrazzaq el Mizalli is the alias used by Ammar Saeefi, emir of the Fifth Region in Algeria for the GSPC. Abassi Madani was a leader in the FIS, an Algerian terrorist group with close ties to the GSPC. El Para is the GSPC commander for the Eastern Region of Algeria.

The overall matrix of groups versus persons for this collection provides similar association information between many other individuals and terrorist groups.

The generality of the LSI representation allows great flexibility in presentation of associations. Any entity type can be compared to any other entity type. Moreover, the rows and columns do not have to correspond to entities. For a given LSI space, a representation vector can be created that corresponds to any arbitrary block of text. Thus, conceptual descriptions can be used as objects for comparison.

As an example of this, Table 4 shows data extracted from one row of a matrix of terrorist groups versus names. In this case, the representation vectors for the terrorist groups were not those associated with their entity names (as in Tables 1 and 2), but rather vectors associated with short (six to eight paragraph) descriptions of the groups. The table shows the names associated with the vectors having the ten largest cosines[**] in the row corresponding to the GSPC description.[††]

The textual description of the GSPC used here focuses on the European operations of the GSPC. Accordingly, the most closely associated individuals in this case tend to be members of the GSPC who were active in Europe. Although the cosines are lower than those generated using the single term GSPC in Table 3, the individuals in Table 4 are even more tightly connected to the GSPC.

Nine of the ten individuals listed in Table 4 are known to be members of the GSPC and the other is the spiritual leader for the GSPC in Europe. During this time period, Abou Doha was the key GSPC leader in London. Rabah Kadri was his deputy. Djamel Beghal was the leader of a GSPC cell, closely supported by Kamel Daoudi. Abu Qutada was appointed by Osama bin Laden to be the spiritual leader for the GSPC in Europe. Hassan Hattab was the founder of the GSPC. Mabrouk Echiker, Laurent Mourad, Meroine Berrahal, and Yacine Akhnouche were all members of the GSPC cell in Frankfurt. The very strong associations of these ten individuals with the GSPC demonstrate the utility of using textual descriptions to define objects for comparison. In general, a textual

[**] Two of the highest cosines in the row corresponded to alternative spellings of Abou Doha and Abou Qutada. They were ignored in creating Table 4.

[††] The textual description of the GSPC was taken from the Center for Defense Information report: In the Spotlight: the Salafist Group for Call and Combat (GSPC), from the CDI website at http://www.cdi.org/terrorism/gspc.cfm

description allows more precise definition of an interest than a single term.

Table 4: Ten names most closely associated with textual description of the GSPC

	Name	Cosine
1	Abou Doha	.5942
2	Djamel Beghal	.5794
3	Abou Qutada	.5623
4	Rabah Kadri	.5593
5	Kamel Daoudi	.5481
6	Mabrouk Echiker	.5442
7	Hassan Hattab	.5410
8	Laurent Mourad	.5371
9	Meroine Berrahal	.5371
10	Yacine Akhnouche	.5371

Textual items of interest for counterterrorism analysis include descriptions of target classes and of potential attack scenarios. In the latter case, the method presented here could be used to continually screen newly acquired information. Receipt of information consistent with a hypothetical or reported attack scenario could automatically trigger an alert to an analyst.

4.4 Entity-entity comparisons – names versus names

Another presentation of interest is person versus person. This display facilitates identification of subgroups. Some of the cosines in this matrix are quite high – implying close association. These implied associations correlate very well with actual associations. For example, the cosine of the angle between the representation vectors for Abou Doha and Abu Qutada is .9347. During the time period covered by the articles, Abou Doha was the senior operational commander for the GSPC in Europe. Abou Qutada was the senior spiritual leader for the group in Europe. The two were close collaborators.

The cosine of the angle between the vectors representing Abou Doha and Rabah Kadri is .8908. Kadri was Doha's deputy in this time frame. Similarly, the cosine of the angle between the representation vectors for Djamel Beghal and Kamel Daoudi is .8991. Daoudi was the principal lieutenant of Beghal in this time period.

A particularly interesting aspect of the person versus person display is that true names and aliases for the same individual typically are closely associated. For example, in this collection, the vector representing the name Ammar Saeefi has a very high cosine value with the vector for the name Abdelrazzaq El Mizalli (cosine = .9675). As noted above, Abdelrazzaq El Mizalli is an alias used by Ammar Saeefi. This coalescence is a very frequent occurrence, and one which can be used to great advantage in counterterrorism analysis.

4.5 Creating and refining entity lists

The final element of this experiment was demonstration of the utility of the LSI representation for creating and refining entity lists. In portions of this experiment, a list of terrorist groups was used as an auxiliary data set. Cross-correlation of this list with the entities identified by LingPipe as organizations yielded a list of terrorist groups that are referred to in the test collection. This list was used in the creation of Tables 1 and 2 in this paper. In general, it is desirable to minimize reliance on auxiliary data sets. Such data sets require user effort to create and maintain and nearly always are incomplete.

A given list of entities of a specific type can be extended in a straightforward manner. This is accomplished through comparison of the representation vectors for the specific entities in the list with those for entities of an appropriate general type that are identified by the entity extraction software. For example, an initial list of terrorist groups can be extended through comparison with entities tagged as organizations. This was demonstrated in this experiment for the set of 14 terrorist groups constituting the rows in Tables 1 and 2. This extension was carried out as follows:

- The 14 representation vectors for the terrorist group entities in Tables 1 and 2 were combined to form a single representation vector in the LSI space. This composite vector can be considered as representing the average of the contexts of the 14 groups.
- The composite vector was compared to the set of entities extracted by the LingPipe software that were tagged as organizations.
- Named entities with high cosine values in this comparison were considered for extension of the list of terrorist groups.

This technique worked very well for this collection. Table 5 shows the ten entities with the highest cosines in this comparison. All ten of these organizations are terrorist groups or elements thereof.

Table 5: Organization entities most similar to composite terrorist group representation vector

	Entity name	Cosine
1	Al-Qassam Brigades	.6593
2	Al-Quds Brigades	.5868
3	Abu-Ali Mustafa Brigades	.5840
4	Democratic Front for the Liberation of Palestine	.5797
5	Fatah Movement	.5335
6	Islamic Jihad General Congress	.5332
7	Moro Islamic Liberation Front	.5166
8	Elite Force 17	.5163
9	Palestinian Arab Front	.5125
10	Army of National Liberation	.4996

This is a general technique that can be used to extend lists of specific entity types. It also can be used to partition entities which are tagged by an entity extractor as to general category into much more finely divided subcategories. Such extension is context dependent. It is most useful in relatively large and content-rich collections.

The approach described above extends entity lists based on proximity to positive examples. In an analogous manner, lists can be pruned based on proximity of entities to negative examples. Moreover, both approaches may be used iteratively.

As noted above, a representation vector can be created in an LSI space that corresponds to any arbitrary block of text. Such a vector represents an average of all of the contexts in which the terms of the text block occur. This allows the creation of entity lists based on textual descriptions. For example, one could start with a textual description of what the user means by a terrorist group. The representation vector for that block of text then could be cross-correlated with the list of entities tagged as organizations by the entity extractor.

Again, this is a general technique. It could be employed to generate candidate entity lists for any concepts that are amenable to contextual comparisons. These candidate lists could be used in either of two ways. If the user were so inclined, he or she could review such a list and prune it of items that they do not wish to include. As an alternative, the user might specify a threshold cosine value and automatically accept all entities with cosines above this value. It should be emphasized that these lists do not need to be completely correct in order to be of utility in the types of analyses presented here. In fact, for reasonable values of cosine threshold, such lists can be more reliable than the typical results obtained using entity extraction packages.

5 Conclusion

The results of this experiment are very encouraging. They demonstrate that the technique employed is capable of identifying numerous associations of interest in large text collections in a completely automated fashion. The approach could be used to provide a rapid overview of large volumes of text. In a dynamic collection environment it could be used to automatically alert users when text has been acquired that contains information of interest.

The approach described here has a number of important characteristics that make it particularly well-suited for intelligence analysis applications:

- It is independent of topic, genre, or language.
- Identified relationships represent the aggregate implications of high-order associations. For this reason, the approach is capable of identifying quite subtle relationships.
- The user has complete flexibility in specifying items of interest. Any point in the LSI representation space can be used as a basis for defining a row or column in the comparison matrices. In particular, a textual description of an entity or concept of interest may be chosen as a definition of an item of interest. That description does not have to have been derived from the documents indexed. There need only be some minimal degree of overlap in terminology between the description and the aggregate terminology employed in the documents used to generate the representation space.
- The technique has significant utility in identifying possible use of aliases.
- There is no need for predefined auxiliary structures such as taxonomies or ontologies. Nor is there need for any linguistic analysis, other than that contained in the entity extraction software.
- The technique can be used in a cross-lingual fashion. The LSI technique is capable of

appropriately representing documents and terms in multiple languages in a single representation space [7]. In such a space, analytic activities that are carried out take advantage of all of the relationship information contained in all of the documents indexed. Thus, this approach can detect the presence of interesting information in collections of foreign-language material without the need to translate any of those documents.[‡‡]

- The LSI technique has been shown to be highly resistant to noise, making it particularly desirable for working with text derived from OCR or speech-to-text conversion operations [13]. Although such noise will hinder entity extraction processes, the results shown here demonstrate that the approach works well even in the absence of highly accurate entity extraction.

This is a work in progress. The LingPipe software employed in the experiment generated a fairly large number of errors of omission and commission in recognizing occurrences of entities. Moreover, it generated a considerable number of incorrect tags.[§§] Examination of these errors indicates that many of them involved specific entities relevant to the examples presented here. The results would have been improved if a more accurate entity extraction capability had been available. Given the level of errors produced by LingPipe, the actual results obtained in the experiment are surprisingly good.

Results also could have been improved by reconciling name variants. A number of techniques are available that would have allowed most of the observed variants to be coalesced automatically [8].

Future efforts will include use of more accurate entity extraction software, reconciliation of name variants, and parametric analysis of thresholds for generating and extending entity lists.

The author would like to thank Martin Hobson, who carried out the entity extraction and LSI indexing operations to generate the representation space used in this study.

References

[1] R. B. Bradford, *Efficient Discovery of New Information in Large Text Databases*, IEEE International Conference on Intelligence and Security Informatics, Atlanta, Georgia, LNCS Vol. 3495, Springer, (2005), pp. 374-380.

[2] S. Dennis, et al, *Introduction to Latent Semantic Analysis,* Slides from tutorial given at the 25th Annual Meeting of the Cognitive Science Society, Boston Massachusetts (2003), lsa.colorado.edu/~quesadaj/pdf/LSATutorial.pdf.

[3] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, L. Beck, *Improving Information Retrieval with Latent Semantic Indexing*, in: Proceedings of the 51st Annual Meeting of the American Society for Information Science, (1988), pp. 36-40.

[4] S. T. Dumais, *Latent Semantic Analysis*, in: Annual Review of Information Science and Technology, Vol. 38. Information Today Inc., Medford, New Jersey (2004), pp.189-230.

[5] A. Kontostathis, W. Pottenger, *Detecting Patterns in the LSI Term-Term Matrix*, Technical Report LU-CSE-02-010, Department of Computer Science and Engineering, Lehigh University, (2002).

[6] A. Kontostathis, W. M. Pottenger, *A Mathematical View of Latent Semantic Indexing: Tracing Term Co-occurrences*, Technical Report LU-CSE-02-006, Department of Computer Science and Engineering, Lehigh University, (2002).

[7] T. Landauer, M. Littman, *Fully Automatic Cross-language Document Retrieval Using Latent Semantic Indexing*, in: Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research (1990), pp. 31-38.

[8] F. Patman, P. Thompson, *Names: A New frontier in Text Mining*, in: First NSF/NIJ Symposium on Intelligence and Security Informatics, Tucson, Arizona, Lecture Notes in Computer Science, Vol. 2665, Springer (2003), pp. 27-38.

[9] D. Skillicorn, *Clusters Within Clusters: SVD and Counterterrorism*, Queen's University School of Computing Technical Report # 2004-363, March 2003.

[10] D. Skillicorn, *Applying Matrix Decompositions to Counterterrorism*, Queen's

[‡‡] An appropriate entity extraction capability would have to be applied for pre-processing documents in each of the languages in the collection.

[§§] This is not meant to be a criticism of LingPipe. Entity extraction is a very difficult problem. Even the most expensive commercial entity extraction software packages make many mistakes in identifying and tagging entities.

University School of Computing Technical Report # 2004-484, May 2004.

[11] D. Skillicorn, *Finding Unusual Correlation Using Matrix Decompositions*, in: Second Symposium on Intelligence and Security Informatics, Tucson, Arizona (2004), pp. 83-99.

[12] D. Skillicorn, *Social Network Analysis via Matrix Decompositions*, Emergent Information Technologies and Enabling Policies for Counter Terrorism, IEEE-Wiley, in press.

[13] A. Zukas, R. J. Price, *Document Categorization Using Latent Semantic Indexing*, in: Proceedings, Symposium on Document Image Understanding Technology (2003), pp. 87-91.