Virtual Weapons for Real Wars: Text Mining for National Security

Alessandro Zanasi

ESRIF-European Security Research and Innovation Forum
University of Bologna Professor
Temis SA Cofounder
a_zanasi@yhaoo.it

Abstract. Since the end of the Cold War, the threat of large scale wars has been substituted by new threats: terrorism, organized crime, trafficking, smuggling, proliferation of weapons of mass destruction. The new criminals, especially the so called "jihadist" terrorists are using the new technologies, as those enhanced by Web2.0, to fight their war. Text mining is the most advanced knowledge management technology which allow intelligence analysts to automatically analyze the content of information rich online data banks, suspected web sites, blogs, emails, chat lines, instant messages and all other digital media detecting links between people and organizations, trends of social and economic actions, topics of interest also if they are "sunk" among terabytes of information.

Keywords: National security, information sharing, text mining.

1 Introduction

After the 9/11 shock, the world of intelligence is reshaping itself, since that the world is requiring a different intelligence: dispersed, not concentrated; open to several sources; sharing its analysis with a variety of partners, without guarding its secrets tightly; open to strong utilization of new information technologies to take profit of the information (often contradictory) explosion (information density doubles every 24 months and its costs are halved every 18 months [1]); open to the contributions of the best experts, also outside government or corporations [2], e.g. through a public-private partnership (PPP or P3: a system in which a government service or private business venture is funded and operated through a partnership of government and one or more private sector companies).

The role of competitive intelligence has assumed great importance not only in the corporate world but also in the government one, largely due to the changing nature of national power. Today the success of foreign policy rests to a significant extent on energy production control, industrial and financial power, and energy production control, industrial and financial power in turn are dependent on science and technology and business factors and on the capacity of detecting key players and their actions.

New terrorists are typically organized in small, widely dispersed units and coordinate their activities on line, obviating the need for central command. Al Qaeda and

E. Corchado et al. (Eds.): CISIS 2008, ASC 53, pp. 53–60, 2009.
springerlink.com

similar groups rely on the Internet to contact potential recruits and donors, sway public opinion, instruct would-be terrorists, pool tactics and knowledge, and organize attacks. This phenomenon has been called Netwar (a form of conflict marked by the use of network forms of organizations and related doctrines, strategies, and technologies) [3]. In many ways, such groups use Internet in the same way that peaceful political organizations do; what makes terrorists' activity threatening is their intent. This approach reflects what the world is experiencing in the last ten years: a paradigm shift from an organization-driven threat architecture (i.e., communities and social activities focused around large companies or organizations) to an individual-centric threat architecture (increased choice and availability of opportunities focused around individual wants and desires). This is a home, local community, and virtual-community-centric societal architecture: the neo-renaissance paradigm. This new lifestyle is due to a growing workforce composed of digitally connected free-agent (e.g. terrorists) able to operate from any location and to be engaged anywhere on the globe [4].

Due to this growing of virtual communities, a strong interest towards the capability of automatic evaluation of communications exchanged inside these communities and of their authors is also growing, directed to profiling activity, to extract authors personal characteristics (useful in investigative actions too).

So, to counter netwar terrorists, intelligence must learn how to monitor their network activity, also online, in the same way it keeps tabs on terrorists in the real world. Doing so will require a realignment of western intelligence and law enforcement agencies, which lag behind terrorist organizations in adopting information technologies [5] and, at least for NSA and FBI, to upgrade their computers to better coordinate intelligence information [6].

The structure of the Internet allows malicious activity to flourish and the perpetrators to remain anonymous.

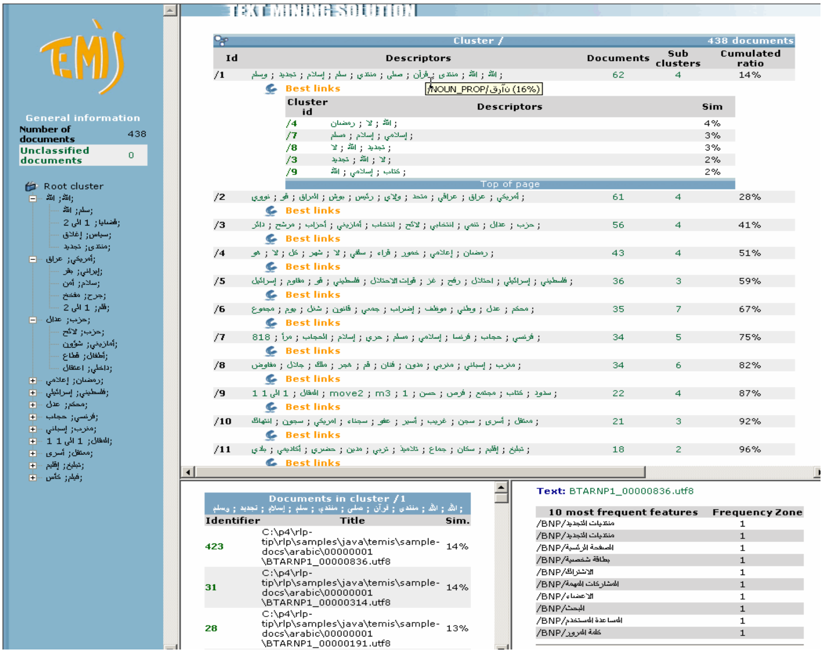
Since it would be nearly impossible to identify and disable every terrorist news forum on the internet given the substantial legal and technical hurdles involved (there are some 4,500 web sites that disseminate the al Qaeda leadership's messages [7]), it would make more sense to leave those web sites online but watch them carefully. These sites offer governments' intelligence unprecedented insight into terrorists' ideology and motivations.

Deciphering these Web sites will require not just Internet savvy but also the ability to read Arabic and understand terrorists' cultural backgrounds-skills most western counterterrorism agencies currently lack [5].

These are the reasons for which the text mining technologies, which allow the reduction of information overload and complexity, analyzing texts also in unknown, exotic languages (Arabic included: the screenshots in the article, as all the other ones, are Temis courtesy), have become so important in the government as in the corporate intelligence world.

For an introduction to text mining technology and its applications to intelligence: [8].

The question to be answered is: once defined the new battle field (i.e. the Web) how to use the available technologies to fight the new criminals and terrorists? A proposed solution is: through an "Internet Center". That is a physical place where to concentrate advanced information technologies (including Web 2.0, machine translation, crawlers, text mining) and human expertise, not only in information technologies but also in online investigations).



General information

Number of documents	438
Unclassified documents	0

Root cluster
- الله; الله;
  - سلم; الله;
  - قضايا; 1 الى 2;
  - سياس; إغلاق;
  - منتدى; تجديد;
- أمريكي; عراق;
  - إيراني; بقر;
  - سلام; أمن;
  - جرح; مفخخ;
  - قلم; 1 الى 2;
- حزب; عدال;
  - حزب; لائح;
  - أمازيغي; شؤون;
  - أطفال; قطاع;
  - داخلي; اعتقال;
- رمضان; إعلامي;
- فلسطيني; إسرائيلي;
- محكم; عدل;
- فرنسي; حجاب;
- مغرب; إسباني;
- المقال; 1 الى 1 1;
- معتقل; أسرى;
- تبليغ; إقليم;
- فيلم; كأس;

Cluster / 438 documents

Id	Descriptors	Documents	Sub clusters	Cumulated ratio
/1	الله ; الله ; منتدى ; قرآن ; صلى ; منتدي ; سلم ; إسلام ; تجديد ; وسلم ;	62	4	14%

Best links /NOUN_PROP/قرآن (16%)

Cluster id	Descriptors	Sim
/4	الله ; لا ; رمضان ;	4%
/7	إسلامي ; إسلام ; مسلم ;	3%
/8	تجديد ; الله ; لا ;	3%
/3	لا ; الله ; تجديد ;	2%
/9	كتاب ; إسلامي ; الله ;	2%

Top of page

Id	Descriptors	Documents	Sub clusters	Cumulated ratio
/2	أمريكي ; عراق ; عراقي ; متحد ; ولاي ; رئيس ; بوش ; الحراق ; قو ; نووي ;	61	4	28%
/3	حزب ; عدال ; تنمي ; انتخابي ; لائح ; انتخاب ; أمازيغي ; أحزاب ; مرشح ; دائر ;	56	4	41%
/4	رمضان ; إعلامي ; خمور ; قراء ; سلفي ; لا ; شهر ; كل ; لا ; هو ;	43	4	51%
/5	فلسطيني ; إسرائيلي ; احتلال ; رفح ; غز ; قوات الاحتلال ; فلسطيني ; قو ; مقاوم ; إسرائيل ;	36	3	59%
/6	محكم ; عدل ; وطني ; موظف ; إضراب ; جمعي ; قانون ; شغل ; يوم ; مجموع ;	35	7	67%
/7	فرنسي ; حجاب ; فرنسا ; إسلامي ; مسلم ; حري ; إسلام ; الحجاب ; مرأ ; 818 ;	34	5	75%
/8	مغرب ; إسباني ; مغربي ; مدون ; فنان ; قم ; هجر ; ملك ; جلال ; مفاوض ;	34	6	82%
/9	سدود ; كتاب ; مجتمع ; فرص ; حسن ; 1 ; m3 ; move2 ; المقال ; 1 الى 1 1 ;	22	4	87%
/10	معتقل ; أسرى ; سجن ; غريب ; أسير ; عفو ; سجناء ; امريكي ; سجون ; انتهاك ;	21	3	92%
/11	تبليغ ; إقليم ; سكان ; جماع ; تلاميذ ; تربي ; مدين ; حضري ; أكاديمي ; بلاي ;	18	2	96%

Best links

Documents in cluster /1
الله ; الله ; منتدى ; قرآن ; صلى ; منتدي ; سلم ; إسلام ; تجديد ; وسلم ;

Identifier	Title	Sim.
423	C:\p4\rlp-tip\rlp\samples\java\temis\sample-docs\arabic\00000001\BTARNP1_00000836.utf8	14%
31	C:\p4\rlp-tip\rlp\samples\java\temis\sample-docs\arabic\00000001\BTARNP1_00000314.utf8	14%
28	C:\p4\rlp-tip\rlp\samples\java\temis\sample-docs\arabic\00000001\BTARNP1_00000191.utf8	13%

Text: BTARNP1_00000836.utf8

10 most frequent features	Frequency Zone
/BNP/منتديات التجديد	1
/BNP/منتديات التجديد	1
/BNP/الصفحة الرئسية	1
/BNP/بطاقة شخصية	1
/BNP/الاشتراك	1
/BNP/المشاركات المهمة	1
/BNP/الاعضاء	1
/BNP/البحث	1
/BNP/المساعدة المستخدم	1
/BNP/كلمة المرور	1

Fig. 1. Text mining analysis of Arabic texts

We present here the scenarios into which these technologies, especially those regarding text mining are utilized, with some real cases.

2 New Challenges to the Market State

The information revolution is the key enabler of economic globalization.

The age of information is also the age of emergence of the so called *market-state* [9] which maximizes the opportunities of its people, facing lethal security challenges which dramatically change the roles of government and of private actors and of intelligence.

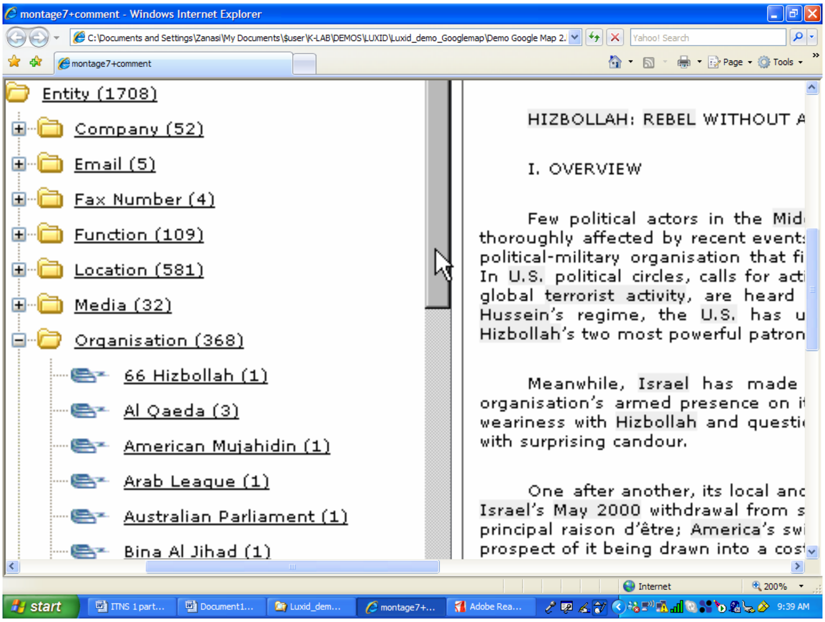
Currently governments power is being challenged from both above (international commerce, which erodes what used to be thought of as aspects of national sovereignty) and below (terrorist and organized crime challenge the state power from beneath, by trying to compel states to acquiesce or by eluding the control of states).

Tackling these new challenges is the role of the new government intelligence.

From the end of Cold War there is general agreement about the nature of the threats that posed a challenge to the intelligence community: drugs, organized crime, and proliferation of conventional and unconventional weapons, terrorism, financial crimes. All these threats aiming for violence, not victory, may be tackled through the help of technologies as micro-robots, bio-sniffers, and sticky electronics. Information technology, applied to open sources analysis, is a support in intelligence activities directed to prevent these threats [10] and to assure homeland protection [11].

Since 2001 several public initiatives involving data and text mining appeared in USA and Europe.

All of them shared the same conviction: *information is the best arm against the asymmetric threats*.


montage7+comment - Windows Internet Explorer
C:\Documents and Settings\Zanasi\My Documents\$user\K-LAB\DEMOS\LUXID\Luxid_demo_Googlemap\Demo Google Map 2.
Yahoo! Search
montage7+comment
Page
Tools
Internet
200%
start
ITNS 1 part...
Document1...
Luxid_dem...
montage7+...
Adobe Rea...
9:39 AM

Entity (1708)
Company (52)
Email (5)
Fax Number (4)
Function (109)
Location (581)
Media (32)
Organisation (368)
66 Hizbollah (1)
Al Qaeda (3)
American Mujahidin (1)
Arab League (1)
Australian Parliament (1)
Bina Al Jihad (1)

HIZBOLLAH: REBEL WITHOUT A

I. OVERVIEW

Few political actors in the Mid
thoroughly affected by recent event
political-military organisation that fi
In U.S. political circles, calls for acti
global terrorist activity, are heard
Hussein's regime, the U.S. has u
Hizbollah's two most powerful patron

Meanwhile, Israel has made
organisation's armed presence on i
weariness with Hizbollah and questi
with surprising candour.

One after another, its local and
Israel's May 2000 withdrawal from s
principal raison d'être; America's swi
prospect of it being drawn into a cost

Fig. 2. The left panel allows us to highlight all the terms which appear in the collected documents and into which we are interested in (eg: Hezbollah). After clicking on the term, in the right column appear the related documents.

3 The Web as a Source of Information

Until some years ago the law enforcement officers were interested only in retrieving data coming from web sites and online databanks (collections of information, available online, dedicated to press surveys, patents, scientific articles, specific topics, commercial data). Now they are interested in analyzing the data coming from internet seen as a way of communication: e-mails, chat rooms, forums, newsgroups, blogs (obtained, of course, after being assured that data privacy rules have been safeguarded).

Of course, this nearly unending stream of new information, especially regarding communications, also in exotic languages, created not only an opportunity but also a new problem. The information data is too large to be analyzed by human beings and the languages in which this data are written are very unknown to the analysts.

Luckily these problems, created by technologies, may be solved thanks to other information technologies.

4 Text Mining

The basic text mining technique is Information Extraction consisting in linguistic processing, where semantic models are defined according to the user requirements, allowing the user to extract the principal topics of interest to him. These semantic models are contained in specific ontologies, engineering artefacts which contain a specific vocabulary used to describe a certain reality, plus a set of explicit assumptions regarding the intended meaning of the vocabulary words.

This technique allows, for example, the extraction of organization and people names, email addresses, bank account, phone and fax numbers as they appear in the data set. For example, once defined a technology or a political group, we can quickly obtain the list of organizations working with that technology or the journalists supporting that opinion or the key players for that political group.

5 Virtual Communities Monitoring

A virtual community, whose blog and chats are typical examples, are communities of people sharing and communicating common interests, ideas, and feelings over the Internet or other collaborative networks. The possible inventor of this term was Howard Rheingold, who defines virtual communities as social aggregations that emerge from the Internet when enough people carry on public discussions long enough and with sufficient human feeling to form webs of personal relationships in cyberspace [12].

Most community members need to interact with a single individual in a one-to-one conversation or participate and collaborate in idea development via threaded conversations with multiple people or groups.

This type of data is, clearly, an exceptional source to be mined [19].

6 Accepting the Challenges to National Security

6.1 What We Need

It is difficult for government intelligence to counter the threat that terrorists pose.

To fight them we need solutions able to detect their names in the communications, to detect their financial movements, to recognize the real authors of anonymous documents, to put in evidence connections inside social structures, to track individuals through collecting as much information about them as possible and using computer algorithms and human analysis to detect potential activity.

6.2 Names and Relationships Detection

New terrorist groups rise each week, new terrorists each day. Their names, often written in a different alphabet, are difficult to be caught and checked against other names already present in the databases. Text mining technology allows their detection, also with their connections to other groups or people.

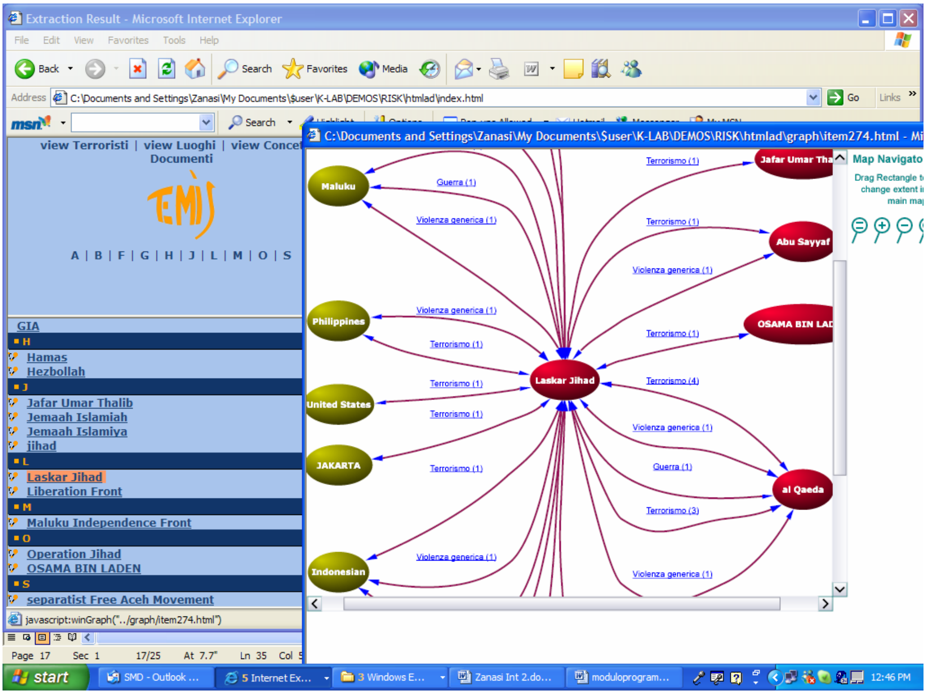


Fig. 3. Extraction of names with detection of connection to suspect terrorist names and the reason of this connection

6.3 Money Laundering

Text mining is used in detecting anomalies in the fund transfer request process and in the automatic population of black lists.

6.4 Insider Trading

To detect insider trading it is necessary to track the stock trading activity for every publicly traded company, plot it on a time line and compare anomalous peaks to company news: if there is no news to spur a trading peak, that is a suspected insider trading.

To perform this analysis it is necessary to extract the necessary elements (names of officers and events, separated by category) from news text and then correlate them with the structured data coming from stock trading activity [13].

6.5 Defining Anonymous Terrorist Authorship

Frequently the only traces available after a terrorist attack are the emails or the communications claiming the act. The analyst must analyze the style, the concepts and feelings [14] expressed in a communication to establish connections and patterns between documents [15], comparing them with documents coming from known authors: famous attackers (Unabomber was the most famous one) were precisely described, before being really detected, using this type of analysis.

6.6 Digital Signatures

Human beings are habit beings and have some personal characteristics (more than 1000 “style markers” have been quoted in literature) that are inclined to persist, useful in profiling the authors of the texts.

6.7 Lobby Detection

Analyzing connections, similarities and general patterns in public declarations and/or statements of different people allows the recognition of unexpected associations («lobbies») of authors (as journalists, interest groups, newspapers, media groups, politicians) detecting whom, among them, is practically forming an alliance.

6.8 Monitoring of Specific Areas/Sectors

In business there are several examples of successful solutions applied to competitive intelligence. E.g. Unilever, text mining patents discovered that a competitor was planning new activities in Brazil which really took place a year later [16].

Telecom Italia, discovered that a competitor (NEC-Nippon Electric Company) was going to launch new services in multimedia [16].

Total (F), mines Factiva and Lexis-Nexis databases to detect geopolitical and technical information.

6.9 Chat Lines, Blogs and Other Open Sources Analysis

The first enemy of intelligence activity is the “avalanche” of information that daily the analysts must retrieve, read, filter and summarize. The Al Qaeda terrorists declared to interact among them through chat lines to avoid being intercepted [17]: interception and analysis of chat lines content is anyway possible and frequently done in commercial situations [18], [19].

Using different text mining techniques it is possible to identify the context of the communication and the relationships among documents detecting the references to the interesting topics, how they are treated and what impression they create in the reader [20].

6.10 Social Network Links Detection

“Social structure” has long been an important concept in sociology. Network analysis is a recent set of methods for the systematic study of social structure and offers a new standpoint from which to judge social structures [21].

Text mining is giving an important help in detection of social network hidden inside large volumes of text also detecting the simultaneous appearance of entities (names, events and concepts) measuring their distance *(proximity*).

References

1. Lisse, W.: The Economics of Information and the Internet. Competitive Intelligence Review 9(4) (1998)
2. Treverton, G.F.: Reshaping National Intelligence in an Age of Information. Cambridge University Press, Cambridge (2001)
3. Ronfeldt, D., Arquilla, J.: The Advent of Netwar –Rand Corporation (1996)
4. Goldfinger, C.: Travail et hors Travail: vers une societe fluide. In: Jacob, O. (ed.) (1998)
5. Kohlmann, E.: The Real Online Terrorist Threat – Foreign Affairs (September/October 2006)
6. Mueller, J.: Is There Still a Terrorist Threat? – Foreign Affairs (September/ October 2006)
7. Riedel, B.: Al Qaeda Strikes Back – Foreign Affairs (May/June 2007)
8. Zanasi, A. (ed.): Text Mining and its Applications to Intelligence, CRM and Knowledge Management. WIT Press, Southampton (2007)
9. Bobbitt, P.: The Shield of Achilles: War, Peace, and the Course of History, Knopf (2002)
10. Zanasi, A.: New forms of war, new forms of Intelligence: Text Mining. In: ITNS Conference, Riyadh (2007)
11. Steinberg, J.: In Protecting the Homeland 2006/2007 - The Brookings Institution (2006)
12. Rheingold, H.: The Virtual Community. MIT Press, Cambridge (2000)
13. Feldman, S.: Insider Trading and More, IDC Report, Doc#28651 (December 2002)
14. de Laat, M.: Network and content analysis in an online community discourse. University of Nijmegen (2002)
15. Benedetti, A.: Il linguaggio delle nuove Brigate Rosse, Erga Edizioni (2002)
16. Zanasi, A.: Competitive Intelligence Thru Data Mining Public Sources - Competitive Intelligence Review, vol. 9(1). John Wiley & Sons, Inc., Chichester (1998)
17. The Other War, The Economist March 26 (2003)
18. Campbell, D.: - World under Watch, Interception Capabilities in the 21st Century – ZDNet.co (2001) (updated version of Interception Capabilities 2000, A report to European Parlement - 1999)
19. Zanasi, A.: Email, chatlines, newsgroups: a continuous opinion surveys source thanks to text mining. In: Excellence in Int'l Research 2003 - ESOMAR (Nl) (2003)
20. Jones, C.W.: Online Impression Management. University of California paper (July 2005)
21. Degenne, A., Forse, M.: Introducing Social Networks. Sage Publications, London (1999)