Text Mining для приложений национальной безопасности

Алыгулиев Рамиз

Институт Информационных Технологий НАН Азербайджана
r.aliguliyev@gmail.com

***Аннотация*— В статье дана краткая информация о целях, задачах и областях применения технологии Text Mining. Анализирована роль этой технологии в системах национальной безопасности и указаны перспективные направления исследований в данной области.**

Ключевые слова— Text Mining, национальная безопасность

I. ВВЕДЕНИЕ

После холодной войны, угроза крупномасштабных войн заменена новыми угрозами, такими как терроризм, организованная преступность, торговля людьми, контрабанда, распространение оружия массового уничтожения. Новые преступники, особенно так называемые «джихадские» террористы используют новые Web технологии, в результате этого бороться с ними становится очень трудно.

Следует отметить, что важность национальной безопасности значительно возросла после террористических атак 11 сентября 2001 года, в результате которых погибло более 3000 невинных людей. Центральное разведывательное управление, Федеральное бюро расследований США и органы национальной безопасности других стран стали активно собирать ключевую информацию от отечественных и иностранных разведок, чтобы предотвратить такие нападения в будущем.

В последние годы наблюдается растущий спрос к применению технологии анализа текста (Text Mining) в системах национальной безопасности (СНБ).

Text Mining является самой передовой технологией управления знаниями, которая позволяет аналитиков разведки автоматически проанализировать содержание информационно богатых онлайн-банков данных, подозреваемых веб-сайтов, блогов, электронной почты, чата мгновенных сообщений и других цифровых связей между людьми и организациями. Например, для выявления незаконной финансовой деятельности корпорации анализируют большую коллекцию электронной почты; аналитики по борьбе с терроризмом просматривают большое количество новостных статей для выявления информации, относящейся к потенциальным угрозам и т.д.

Одним из таких приложений Text Mining в СНБ является глобальная система радиоэлектронной разведки сети и анализа сигналов – «Эшелон» [6]. Технология Text Mining использована системой «Эшелон» для обнаружения террористических заговоров, планов наркоторговцев, политической и дипломатической разведки.

Официальная история «Эшелона» начинается в 1947 году, когда между США и Англией было заключено секретное соглашение «UK-USA Agreement» (UKUSA Signals Intelligence Agreement – Соглашение о радиотехнической разведывательной деятельности Великобритания - США), по которому эти страны объединяли свои технические и человеческие ресурсы в сфере глобального электронного шпионажа. Базой для «Эшелона» послужили мощные подразделения технической разведки, созданные в годы второй мировой войны спецслужбами США и Великобритании. Чуть позже к США и Англии присоединились Канада, Австралия и Новая Зеландия. Руководители радиоразведок стран «пятерки» ежегодно собирались вместе, чтобы обсудить вопросы планирования и координации деятельности по направлениям глобальной разведки. В работе альянса также участвуют Германия, Япония, Республика Корея, Турция и Норвегия, а в последние годы и страны-участники НАТО.

«Эшелон» имеет возможность перехвата и анализа телефонных переговоров, факсов, электронных писем и других информационных потоков по всему миру путём подключения к каналам связи, таким как спутниковая связь, телефонная сеть общего пользования, СВЧ-излучение, оптоволоконные соединения и др. [13].

В современном мире организации используют Text Mining для извлечения информации с целью более эффективного управления знаниями. Трагическое событие 9 сентября 2011 года побудило правительство США увеличить ресурсы, чтобы обеспечить гарантийную безопасность страны. Поэтому развертывание технологии Text Mining стало жизненно важным в СНБ. Кроме того, безопасность компьютерных информационных систем и коммуникационной инфраструктуры на сегодняшний день является одним из национальных приоритетов стран.

В поисках новой технологической разработки проведено множество исследований в области Text Mining, однако мало внимания уделялось ее использованию в СНБ. В попытке заполнить пробел, настоящая статья стремится пролить свет на роль Text Mining в СНБ.

II. ЦЕЛИ И ЗАДАЧИ ТЕХНОЛОГИИ TEXT MINING

Технология Text Mining представляет собой одну из разновидностей методов Data Mining [8] и подразумевает процессы извлечения знаний и высококачественной информации из текстовых массивов. Извлечение знаний из текстов – это процесс обнаружения новых, потенциально полезных и понятных шаблонов в неструктурированных текстовых данных – в наборе документов, представляющих собой логически объединенный текст без каких-либо ограничений на его структуру [8, 9, 10]:

- веб-страницы,
- электронная почта,
- нормативные документы,
- мобильные текстовые сообщения и т.д.

Такая технология глубинного анализа текстов способна «просеивать» большие объемы неструктурированной информации и выявлять из них только самое значимое, чтобы человеку не приходилось самому тратить время на добычу ценных знаний «вручную» [10].

По сути, Text Mining – это набор лингвистических, статистических методов, а также алгоритмов машинного обучения, которые способны моделировать и структурировать информационный контент и текстовые источники в целях бизнес-аналитики, анализа данных, исследований.

Text Mining состоит из следующих этапов [10]:

1. Поиск информации.
2. Предварительная обработка документов.
3. Извлечение информации.
4. Применение методов Text Mining.
5. Интерпретация результатов.

Основные задачи Text Mining [10]:

- ***Классификация*** **–** определение для каждого документа одной и нескольких заранее заданных категорий, к которой этот документ относится.
- ***Кластеризация*** **–** автоматическое выявление групп семантически похожих документов среди заданного фиксированного множества.
- ***Автоматическое реферирование*** **–** позволяет сократить текст, сохраняя его смысл.
- ***Извлечение ключевых понятий*** **–** идентификация фактов и отношений в тексте.
- ***Навигация по тексту*** **–** позволяет перемещаться по документам относительно тем и значимых терминов.
- ***Поиск ассоциаций*** – идентификация ассоциативных отношений между ключевыми понятиями.

Технология Text Mining может быть применена в любой области, где существуют текстовые документы. Например, она используется для управления знаниями в различных направлениях и областях – это может быть использование в правительственных, исследовательских, корпоративных целях. Ниже перечислены наиболее типичные отрасли использования этой технологии [10]:

- Корпоративная бизнес-аналитика и Data Mining, корпоративная разведка.
- Делопроизводство и электронные исследования.
- Национальная безопасность и разведка.
- Научные исследования, особенно в естественнонаучной сфере.
- Смысловой анализ.
- Исследование естественных языков и семантики.
- Издательское дело.
- Автоматизированное размещение рекламы.
- Поиск информации и предоставление доступа к ней.
- Мониторинг социальных медиа.

В следующем разделе описаны перспективные направления приложений технологии Text Mining в СНБ.

III. ПЕРСПЕКТИВНЫЕ НАПРАВЛЕНИЯ ПРИЛОЖЕНИЙ TEXT MINING В НАЦИОНАЛЬНОЙ БЕЗОПАСНОСТИ

В настоящее время государственной разведке трудно бороться с угрозами, которые представляют террористы. В борьбе с терроризмом необходимы решения, способные: 1) обнаружить имена террористов в коммуникации, их финансовых движений, распознавать реальных авторов анонимных документов; 2) использовать в качестве доказательства связи в социальных структурах; 3) следить за сомнительными людьми путем сбора и анализа информации о них; 4) использовать компьютерные алгоритмы для выявления потенциальной активности террористов [1-5, 7, 9, 11, 12, 14, 15, 16].

Ниже перечислены наиболее перспективные направления использования Text Mining в СНБ [14, 15].

Обнаружение подозрительных лиц и их скрытых отношений. Новые террористические группы появляются каждую неделю, а новые террористы каждый день. Их имена, часто записанные в другом алфавите, трудно поймать и проверить, напротив других имен, которые уже присутствуют в базах данных. В таком случае технология Text Mining позволяет обнаружить их имена, а также их связи с другими группами или людьми [2, 3, 4, 5, 9].

Инсайдерская торговля. Для выявления инсайдерской торговли необходимо отслеживать активность биржевой торговли для каждой общественной компании, построить ее на временной шкале и сравнить аномальные пики с новостями компании: имеются какие-то новости, стимулирующие торговые пики, то есть подозрительная информация в инсайдерской торговле. Для выполнения этого анализа необходимо извлечь информацию из текстовых новостей, а затем соотнести их со структурированными данными, поступающими с акции торговой деятельности.

Обнаружение лобби. Анализ связи, сходства и общие закономерности в общественных декларациях и/или заявлениях разных людей позволяют распознавать неожиданные ассоциации («лобби»).

Анализ чатов, блогов и других открытых источников. Первым врагом разведывательной

деятельности является «лавина» информации, которую аналитики должны ежедневно извлекать, читать, фильтровать и резюмировать. Террористы Аль-Каиды осуществляют взаимодействие через чат, чтобы избежать перехвата. Как известно, в настоящее время перехват и анализ содержания чата в любом случае возможен и часто делается в коммерческих решениях. Использованием различных методов Text Mining можно определить контекст связи и отношения между документами, обнаружить ссылки к интересным темам и можно определить, как с ними обращаются, и какое впечатление они создают у читателя [2, 3, 4].

Борьба с отмыванием денег. Text Mining используется для выявления аномалий в финансовых операциях и в автоматической популяции черных списков.

Мониторинг отдельных секторов. В бизнесе можно привести несколько примеров, в которых успешно применялись технологии Text Mining в конкурентной разведке. Например, Unilever – патент по технологии Text Mining обнаружил, что конкурент планирует новые мероприятия в Бразилии, которые действительно имели место через год. Telecom Italia, обнаружил, что конкурент (NEC-Nippon Electric Company) собирается запустить новые услуги в области мультимедиа. Total (F) анализирует базы данных Factiva и Lexis-Nexis для определения геополитической и технической информации [14].

Обнаружение социальных сетей. «Социальная структура» давнее и важное понятие в социологии. Анализ сети – это набор методов для систематического изучения социальной структуры, предлагающий новую точку зрения, в результате которой можно судить о социальных структурах.

Text Mining играет важную роль в обнаружении социальных сетей, скрытых в больших объемах текстов, а также в обнаружении совместного появления имен и событий, измерением их близости.

Обнаружение анонимных террористических авторств. Следами часто после террористических атак являются электронные письма. Аналитик должен анализировать стиль, понятия и чувства, выраженные в сообщениях, установить связи и закономерности между документами, сравнивая их с документами, поступающими от известных авторов [1, 15, 16].

ЗАКЛЮЧЕНИЕ

Последние годы терроризм вызывает шоковый эффект во всем мире. Было установлено, что террористы используют современные технологии коммуникации в качестве среды для передачи информации.

Для предотвращения намерений террористов необходимо иметь такие системы, которые могли бы идентифицировать информацию, которой они обмениваются. Правительственные агентства вкладывают значительные средства в наблюдение за всеми видами связи, такими, как электронная почта, чаты, блоги, социальные сети, форумы и т.д. Эти виды связи в настоящее время являются самыми распространенными видами коммуникации, которые используются во многих легитимных видах деятельности. К сожалению, такие виды общения используются не в хороших целях, например, при распределении оскорбительных, угрожающих и т.д. материалов. Так как время имеет решающее значение, учитывая масштабы проблемы, традиционные методы не в силах анализировать такие виды материалов. В таком случае, использование новых технологий в области национальной безопасности становится крайне важным вопросом.

Итак, после анализа мы пришли к выводу, что одним из перспективных направлений использования технологии Text Mining является система национальной безопасности.

ЛИТЕРАТУРА

[1] A. Abbasi, H. Chen, "Applying authorship analysis to extremist group web forum messages," *IEEE Intelligent Systems*, no.5, pp.67-75, 2005.
[2] R. M. Alguliev, R. M. Aliguliyev, S. A. Nazirova, "Classification of textual e-mail spam using data mining techniques," *Applied Computational Intelligence and Soft Computing*, vol.2011, Article 416308, 8 pages, 2011.
[3] T. A. Almeida, A. Yamakami, "Advances in spam filtering techniques," *Computational Intelligence for Privacy and Security Studies in Computational Intelligence*, vol. 394, pp.199-214, 2012.
[4] M.W. Berry, M. Browne, "E-mail surveillance using nonnegative matrix factorization," *Computational & Mathematical Organization Theory*, vol.11, no.3, pp.249-264, 2005.
[5] Y. Elovici, B. Shapira, M. Last, A. Kandell, O. Zaafrany, "Using data mining techniques for detecting terror-related activities on the web," Journal of Information Warfare, vol.3, no.1, pp.17-28, 2004.
[6] European Parliament. European Parliament report on ECHELON. http://www.fas.org/irp/program/process/ rapport_echelon_en.pdf, 2001
[7] D. Guthrie, "Unsupervised detection of anomalous text," PhD Thesis, University of Sheffield, 2008.
[8] J. Han, M. Kamber, *Data mining: concepts and techniques*, 2nd edition, Morgan Kaufmann Publishers, 2006.
[9] M. J. H. Lim, "Computational intelligence in email traffic analysis", PhD Dissertation, University of Tasmania, 2008.
[10] I. Mani, M. Maybury, "*Advances in automatic text summarization*," Cambridge: MIT Press, 1999. 442p.
[11] S. Nizamani, N. Memon, U.K. Wiil, and P. Karampelas, "Modeling suspicious email detection using enhanced feature selection," *International Journal of Modeling and Optimization*, vol.2, no.4, pp.371-377, 2012.
[12] G. Wang, H. Chen, H. Atabakhsh, "Automatically detecting deceptive criminal identities," *Communications of the ACM*, vol.47, no.3, pp.71-76, 2004.
[13] http://www.wikipedia.org/
[14] A. Zanasi, "Competitive intelligence through data mining public sources," *Competitive Intelligence Review*, vol.9, no.1, pp.44-54, 1998.
[15] Zanasi, "Virtual weapons for real wars: text mining for national security," in: *Proceedings of the International Workshop on Computational Intelligence in Security for Information Systems. Advances in Soft Computing*, 2009, vol.53, pp.53-60.
[16] R. Zheng, J. Li, H. Chen, Z. Huang, "A framework of authorship identification for online messages: writing style features and classification techniques," *Journal of the American Society for Information Science and Technology*, vol.57, no.3, pp.378-393, 2006.