Инженерный вестник Дона, №4 (2015)
ivdon.ru/ru/magazine/archive/n4y2015/3409

Методика и фреймворк конструирования лингвистических моделей для сетевого мониторинга

В.И. Носко, В.П. Свечкарев, М.Д. Розин

Южный федеральный университет, Ростов-на-Дону

Аннотация: Высочайшая динамика процессов проявления экстремизма задает темп развитию методологии и программного инструментария, способных в режиме реального времени отслеживать распространение информации, в том числе, в социальных сетях, анализировать ее смыслы и посылы и строить прогностические модели развития ситуаций. Представлена методика умного конструирования лингвистических моделей, которые способны учитывать контекст и гибко адаптироваться под предметную область для анализа текстов в социальных сетях в рамках прикладной задачи упреждения информационно-управляемых угроз на основе технологий Data Mining . Выделены основные недостатки использования простого инжиниринга атрибутов (feature engineering) и метода «мешок слов» (bag of words) в задачах классификации текстов. Описан программный интерфейс и возможности фреймворка, в котором эта методика применяется и показано, как данный фреймворк может использоваться для решения задач бизнеса и государства в процессе сбора и анализа публикаций в сети Интернет.
Ключевые слова: обработка естественного языка, лингвистическая модель, машинное обучение, инжиниринга атрибутов, фреймворк обработки текста, классификация текстов, конструктор языковых моделей, морфологический анализ.

Актуальность проблемы сетевого мониторинга определяется спецификой современной глобальной коммуникативной ситуации, отражающей рост активности непрерывно умножающегося сообщества участников сетевого взаимодействия. Особенно актуализируется мониторинг таких взаимодействий в конфликтных ситуациях, когда наблюдаются признаки как скрытой (латентной), так и открытой агрессии и проявлений экстремизма, способных вызвать социальное напряжение. В дальнейшем формируются узлы напряженности, увязанные между собой множеством разнообразных отношений и связей, и, зачастую, имеющих скрытое информационное управление, распространяемое по ключевым информационным ресурсам. Причем утрата внешней очевидности такого рода угроз не делает их менее опасными для национальной безопасности России [1, 2]. В подобной ситуации методологически оправдана реализация противодействия

экстремистской сетевой агрессии, позволяющего в долгосрочной перспективе предотвращать формирование узлов напряженности и среды её поддержки и инициализации [3, 4]. Высочайшая динамика процессов проявления экстремизма задает темп развитию методологии и программного инструментария, способных в режиме реального времени отслеживать распространение информации, в том числе, в социальных сетях, анализировать ее смыслы и посылы и строить прогностические модели развития ситуаций. В условиях, когда непрерывно генерируются всё новые угрозы террористических атак, видоизменяются попытки дестабилизировать ситуацию и совершенствуются механизмы вовлечения в экстремистскую деятельность населения, требуется формирование методологии и научного инструментария исследования принципиально нового уровня. Информационные и интеллектуальные методы мониторинга и оценки структуры и динамики социальных сетей позволяют находить очаги зарождения потоков информации, в том числе конкретных агентов влияния, а также группы таких агентов, действующих в едином смысловом поле. Такой подход позволяет находить скрытые связи и закономерности, модели поведения и потребления культурных и социальных инициатив.

Следует отметить, что методологическая проблема применения анализа взаимодействий в социальных сетях в прикладной задаче упреждения динамичных процессов является достаточно новой. В статье М. Ньюмана [5] дается подробный обзор существующих сетей, включая социальные, информационные, технологические и биологические, выявляются концептуальные сходства и различия между ними. Рассматриваются такие свойства сетей как кластеризация, распределение степеней, длина пути и др. В статье [6] авторы подробно рассматривают концепцию управления сложными сетями, вводят понятие «драйвер», приводят анализ того, каким образом ограниченная группа активных участников сети может влиять на ее

структурные и динамические свойства. В книге [7] предпринята попытка систематизировать математические модели, касающиеся информационного влияния, управления и противоборства в социальных сетях. Далее проведен краткий анализ использования теоретико-игровых моделей для описания взаимодействия пользователей (агентов) в социальной сети (графовая структура данных), а также представлен программный комплекс для динамическиго моделирования интересующих исследователя жизненных сценариев. В [8] рассматриваются вопросы, относящиеся к информационной структуре сетевого пространства, теории сложных сетей, моделям информационного поиска и глубинного анализа текстов, общим закономерностям современных информационных потоков и их моделированию. Несмотря на значительный объём и результаты вышеназванных работ, следует отметить, что в целом проблема сетевого мониторинга во всей полноте и многообразии в научной литературе исследована недостаточно, в частности, именно так обстоит дело с методиками и программными продуктами конструирования поисковых моделей, необходимых для указанного мониторинга.

В настоящей работе предпринята попытка заполнения данного пробела, а именно представляется одна из возможных реализаций фреймворка конструирования лингвистических моделей для сетевого мониторинга на основе методики их умного конструирования, которая позволяет учитывать контекст и гибко адаптироваться под предметную область.

Конструирование лингвистических моделей как одно из средств отображения языковых явлений и процессов, применяется в органическом единстве с другими методами изучения языка [9]. Модельное конструирование выступает как средство углубления познания скрытых механизмов речевой деятельности, его движения от относительно примитивных моделей к более содержательным моделям, полнее

раскрывающим сущность языка. Так, известные недостатки использования простого инжиниринга атрибутов (feature engineering) и метода «мешок слов» (bag of words) в задачах классификации текстов, связанные с принципиальной невозможностью количественного описания качественных данных, динамики и взаимосвязи данных, предполагают привлечение и согласованное использование более сложных и разнообразных методов. Действительно, если атрибуты обеспечивают детальное описание сущностей, то принципиально недопустимо замыкаться на простых или атомарных атрибутах, вынося за скобки рассмотрения составные, многозначные и производные атрибуты. Мешок слов (или Bag of Words) обычно представляется как модель текстов на натуральном языке в виде неупорядоченного набора слов без сведений о связях между ними, что принципиально исключает как качественный анализ, так и анализ структуры распространения и его динамики.

Разработанная методика конструирования моделей представляет собой комбинацию лучших мировых практик в области анализа текста, в частности, в ней применяется иерархический анализ текста, определение именованных сущностей (named entityrecognition), определение части речи (part-of-speech tagging) и морфологические признаки слов (род, число, падеж) а также правила их согласования. В методике применяется GLR-парсер (Generalized Left-to-right Rightmost derivation parser) — расширенный алгоритм LR-парсера, предназначенный для разбора по недетерменированным и неоднозначным грамматикам. Впервые он был описан Масару Томита в 1984 году [10], его также называют «параллельным парсером».

В методике развиты наработки, описанные в [11], а именно, принципиальная ориентация на автоматическое генерирование независимых признаков поиска с возможностью в дальнейшем адаптации и обучения.

В методике используется подход, основанный на архитектурной организации процесса или конвейера с логическим ветвлением, состоящей из подпроцессов – контекстно-зависимых грамматик, каждая из которых получает на вход результат работы предыдущего подпроцесса. Новизна предлагаемой методики связана с тем, что каждый подпроцесс в свою очередь состоит из вложенных подпроцессов, уровень вложенности при этом ничем не ограничен и зависит от исследователя и конкретной прикладной задачи, а также необходимого уровня точности, которого нужно достичь в решении. Оригинальность методики заключается в том, что используется иерархичность при составлении грамматик, что позволяет исследователю в процессе решения прикладной задачи конструировать сложные смысловые атрибуты и задавать правила ранжирования и важности входящих в них более простых атрибутов.

Подпроцессы при этом состоят из грамматик и связей между ними. Важно отметить, что правила, по которым составляются грамматики, подразделяются по уровню абстракции на 4 фундаментальных класса, или уровня:

1. Уровень символов, которые задаются, как правило, регулярными выражениями;
2. Морфологический анализ, уровень слов, которые задаются онтологиями предметной области, словарями синонимов, определяются частями речи (part-of-speech tagging), морфологическими признаками слов (род, число, падеж) и правилами их согласования;
3. Синтаксический анализ, уровень предложения, когда определяется структура и связи между словами на основе роли слов и их частей речи в предложении: подлежащее, сказуемое и т.п.;
4. Смысловой анализ, уровень документа, на котором уже сформированы сложные составные атрибуты (features) с применением определения

именованных сущностей (named entity recognition) и правил извлечения смысловых блоков в документе. На этом уровне применяются логические операции «и», «или», «не» и другие, а также функции агрегации: длина, частотность и другие.

Таким образом, методика позволяет формализовать сложную задачу классификации текстов и разделить ее на меньшие составные задачи, четко предопределяя последовательность и порядок обработки данных. Например, традиционная задача классификации текстов на два класса: позитив/негатив – может быть решена при помощи простого подхода на основе словарей тональности. В результате для каждого текста на входе будет указан класс, к которому он относится: позитив или негатив. Однако такое решение в настоящее время уже не устраивает бизнес и госструктуры. Предложенная методика позволяет повысить качество определения тональности, определяя дополнительно к какому объекту в тексте относится выражаемое мнение, какой это тип объекта, а также какие именно ключевые характеристики объекта описываются во мнении. Методика позволяет конструировать языковые модели для решения задач с нечеткой логикой постановки, например: поиск агитации, экстремистских высказываний, религиозных и иных призывов, поиск информации о конкурентах, анализ языковых трендов и мемов в среде интернет-аудитории [12].

Описанная методика определяет и архитектуру фреймворка, т.е. структуру программной системы, собственно реализующей задачу конструирования лингвистических моделей и облегчающей, в свою очередь, процесс взаимодействия пользователя за счет удобного адаптируемого интерфейса. Такая организация является очень полезной, потому что создается возможность использовать многоразовые конструкты, которые обеспечивают некоторую расширенную функциональность [13]. Базовые шаблоны

конструирования являются довольно примитивными и их очень легко запомнить и использовать далее для наращивания функциональности.

Таким образом, предложенные методика и фреймворк предназначены для мониторинга и оценки структуры и динамики социальных сетей. Описанные методика и фреймворк позволяют выйти на решение актуальной научной задачи мониторинга динамики речевого экстремизма в русскоязычной сетевой коммуникации, результаты мониторинга могут быть использованы для организации противодействия экстремизму, для проактивного управления в условиях сетевой агрессии.

Литература

1. Иванова М.И., Клаус Н.Г., Литвинов С.В., Мощенко И.Н., Носко В.И., Розин М.Д., Свечкарев В.П., Сущий С.Я., Тымчук Д.А., Угольницкий Г.А. Современная практика моделирования этносоциокультурной конфликтности на Юге России /Под ред. М.Д. Розина. Ростов н/Д: СКНЦ ВШ ЮФУ, 2012. 160 с.

2. Розин М.Д., Свечкарев В.П., Конторович С.Д., Литвинов С.В., Носко В.И. Проблемы мониторинга социальных сетей как площадки социальной коммуникации рунета // Научная мысль Кавказа. Междисциплинарные и специальные исследования, 2011, №2. С.65-77.

3. Rozin M.D., Svechkarev V.P., Mochtchenko I.N., Ryabtsev V.N., Suschiy S.Y. Forecast Evaluation of the Social and Political Tensions Potential for the Proactive Countermeasures against Extremism. Asian Social Science; Vol. 11, No. 6; 2015. pp. 214-220.

4. Свечкарев В.П. Технологии проактивного противодействия экстремизму // Инженерный вестник Дона, 2014. №4. URL: http://www.ivdon.ru/ru/magazine/archive/N4y2014/2606

5. Newman M. E. J. The structure and function of complex networks. SIAM REVIEW. Society for Industrial and Applied Mathematics. Vol. 45, No. 2, 2003, pp. 167–256.

6. Yang-Yu Liu, Jean-Jacques Slotine, Albert-László Barabási. Controllability of complex networks. Nature, vol. 473, no. 7346, 2011, pp. 167-173.

7. Губанов Д.А., Новиков Д.А., Чхартишвили А.Г. Социальные сети: модели информационного влияния, управления и противоборства / Под ред. чл-корр. РАН Д.А. Новикова. - М.: Изд-во физ.-мат. литературы, 2010. 228 с.

8. Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы. - М.: Либроком (Editorial URSS), 2009. 264 с.

9. Моделирование языковой деятельности в интеллектуальных системах. Под ред. А.Е. Кибрика и А.С. Нариньяни; с предисловием А.П. Ершова. - М.: Наука, Гл. ред. физ.-мат. лит., 1987. 280 с.

10. Masaru Tomita (1984). LR parsers for natural languages. COLING. 10th International Conference on Computational Linguistics. pp. 354–357

11. Rami Al-Rfou, Bryan Perozzi, Steven Skiena. Polyglot: Distributed Word Representations for Multilingual NLP. Proceedings of the Seventeenth Conference on Computational Natural Language Learning, Sofia, Bulgaria, http://www.aclweb.org/anthology/W13-3520.

12. Носко В.И. Применение теории графов в интеллектуальной методике анализа социальных медиа для мониторинга популярности кандидатов в период предвыборной кампании // Инженерный вестник Дона, 2014. №3. URL: http://ivdon.ru/ru/magazine/archive/n3y2014/2528

13. Рассел Д., Кон Р. Фреймворк – М.: Изд-во Книга по требованию, 2012. 208 с.

References

1. Ivanova M.I., Klaus N.G., Litvinov S.V., Moshhenko I.N., Nosko V.I., Rozin M.D., Svechkarev V.P., Sushhij S.Ja., Tymchuk D.A., Ugol'nickij G.A. Sovremennaya praktika modelirovaniya etnosotsiokul'turnoy konfliktnosti na Yuge Rossii [Modern practice of modeling ethno-socio-cultural conflicts in the South of Russia]. Pod red. M.D. Rozina. Rostov n/D: SKNC VSh YuFU, 2012. 160 p.

2. Rozin M.D., Svechkarev V.P., Kontorovich S.D., Litvinov S.V., Nosko V.I. Nauchnaya mysl' Kavkaza. Mezhdistsiplinarnye i spetsial'nye issledovaniya (Rus), 2011, №2. pp.65-77.

3. Rozin M.D., Svechkarev V.P., Mochtchenko I.N., Ryabtsev V.N., Suschiy S.Y. Forecast Evaluation of the Social and Political Tensions Potential for the Proactive Countermeasures against Extremism. Asian Social Science; Vol. 11, No. 6; 2015. pp. 214-220.

4. Svechkarev V.P. Inženernyj vestnik Dona (Rus), 2014. №4. URL: ivdon.ru/ru/magazine/archive/N4y2014/2606

5. Newman M. E. J. The structure and function of complex networks. SIAM REVIEW. Society for Industrial and Applied Mathematics. Vol. 45, No. 2, 2003, pp. 167–256.

6. Yang-Yu Liu, Jean-Jacques Slotine, Albert-László Barabási. Controllability of complex networks. Nature , vol. 473, no. 7346, 2011, pp. 167-173.

7. Gubanov D.A., Novikov D.A., Chkhartishvili A.G. Sotsial'nye seti: modeli informatsionnogo vliyaniya, upravleniya i protivoborstva [Social networks: models of information influence, control and confrontation]. Pod red. chl-korr. RAN D.A. Novikova. M.: Izd-vo fiz.-mat. literatury, 2010. 228 p.

8. Lande D.V., Snarskiy A.A., Bezsudnov I.V. Internetika: Navigatsiya v slozhnykh setyakh: modeli i algoritmy [Internetika: Navigation in complex networks: models and algorithms]. M.: Librokom (Editorial URSS), 2009. 264 p.

9. Modelirovanie yazykovoy deyatel'nosti v intellektual'nykh sistemakh [Modelling of linguistic activities in intelligence systems]. Pod red. A.E. Kibrika i A.S. Narin'yani; s predisloviem A.P. Ershova. M.: Nauka, Gl. red. fiz.-mat. lit., 1987. 280 p.

10. Masaru Tomita (1984). LR parsers for natural languages. COLING. 10th International Conference on Computational Linguistics. pp. 354–357

11. Rami Al-Rfou, Bryan Perozzi, Steven Skiena. Polyglot: Distributed Word Representations for Multilingual NLP. Proceedings of the Seventeenth Conference on Computational Natural Language Learning, Sofia, Bulgaria, http://www.aclweb.org/anthology/W13-3520.

12. Nosko V.I. Inženernyj vestnik Dona (Rus), 2014. №3. URL: http://ivdon.ru/ru/magazine/archive/n3y2014/2528

13. Rassel D., Kon R. Freymvork [Framework]. M.: Izd-vo Kniga po trebovaniyu, 2012. 208 p.