УДК 004.82

П. В. Мызников

Новосибирский государственный университет
ул. Ляпунова, 4, Новосибирск, 630009, Россия

miznikov72@gmail.com

МОДЕЛИРОВАНИЕ РАССУЖДЕНИЙ НА ОСНОВЕ ПРЕЦЕДЕНТОВ В АВТОМАТИЧЕСКОМ АНАЛИЗЕ НОВОСТНЫХ ТЕКСТОВ

Рассматриваются методы моделирования рассуждений на основе прецедентов в задачах анализа и моделирования распространения сообщений в сети Интернет. Предлагаемый подход позволяет строить последовательности рассуждений о содержании новостного текста. Особое внимание уделено поиску аргументов и контраргументов в похожих текстах для построения позиций нескольких точек зрения на содержание новости.

Ключевые слова: моделирование рассуждений, прецеденты, обработка естественных языков, анализ текста.

Введение

Ресурсы сети Интернет являются эффективными новостными каналами с точки зрения охвата аудитории и скорости распространения информации. С другой стороны, объем данных и их слабая структурированность вызывают проблемы с анализом такой информации и моделированием ее распространения.

Построение модели распространения новостных сообщений способно повлиять на качественное улучшение решения нескольких задач, интересных для распространителей информации, а именно для оценки:

1) охвата аудитории новостного сообщения,
2) вероятности попадания сообщения в определенный новостной источник,
3) степени интереса пользователей к определенной новости.

Однако для выполнения этих задач необходимо иметь процедуру формализации текста новости, чтобы получить объекты, к которым можно применять вычислительные методы. Большинство существующих подходов направлено либо на статистическое представление текстов (TF-IDF, Bag of words), либо на построение синтаксических деревьев. Такие подходы хорошо справляются с кластеризацией текстов, извлечением фактов и другими задачами. Тем не менее они не отображают важную деталь, которая необходима в поставленных нами задачах, а именно интерпретацию текста с разных точек зрения. Люди по-разному воспринимают информацию, учет этой особенности позволит более точно моделировать распространение сообщений в Интернете.

Для моделирования рассуждений используется методология рассуждения на основе прецедентов.

Особенности новостного текста как объекта автоматической обработки

Прежде чем перейти к описанию подхода, выделим важные особенности новостных сообщений как текста.

Мызников П. В. Моделирование рассуждений на основе прецедентов в автоматическом анализе новостных текстов // Вестн. НГУ. Серия: Информационные технологии. 2017. Т. 15, № 2. С. 59–65.

ISSN 1818-7900. Вестник НГУ. Серия: Информационные технологии. 2017. Том 15, № 2

Принцип перевернутой пирамиды

В работах, посвященных лингвистическому анализу новостного текста (см., например, [1; 2]), подчеркивается так называемый принцип перевернутой пирамиды: самая ценная и важная информация сообщается в начале текста, при этом основная информационная нагрузка приходится на первую фразу, которая называется «лид». По мере развертывания текста информационная нагрузка постепенно ослабевает.

Для решения рассматриваемой задачи это означает, что порядковый номер предложения в тексте является важным атрибутом в анализе текста.

Структура новостного текста

Новостной текст имеет определенную структуру. Какие-то его структурные элементы обязательны, какие-то нет; какие-то имеют строго зарезервированное место, какие-то могут располагаться в разных местах.

I. Заголовок (обозначение: T)

II. Лид (обозначение: !). «Новость одним предложением». Размещается в самом начале текста, содержит *главное событие*. Состоит из одного, иногда двух предложений.

a) резюмирующий: в лиде дается краткое изложение события или выделяется главное, что произошло, а сама статья посвящается раскрытию темы, представляет подробности ($!^{r}$);

b) отложенный: такой лид дает меньше информации, но втягивает в чтение; информация последует далее, когда читатель уже увлечен ($!^{d}$);

c) эпизодный: рассказывается некий эпизод, связанный с тем, что будет в статье ($!^{e}$);

d) цитатный: высказывание известного человека ($!^{q}$).

III. Более подробное описание события (обозначение: D).

IV. Фоновая справка (обозначение: b)

a) хронологическая – содержит историю вопроса;

b) контекстуальная – содержит информацию о сопутствующих обстоятельствах.

V. Изложение позиции одной стороны (обозначение: $\rightarrow$)

VI. Изложение позиции другой стороны (обозначение: $\leftarrow$). (Данный пункт является парным с предыдущим. Либо присутствуют в тексте вместе, либо не присутствует ни один.)

VII. Описание реакции общественности (обозначение: r)

VIII. Прогноз эксперта (обозначение: p).

IX. Комментарий (обозначение:). Содержит прямое или опосредованное мнение автора статьи к главному событию.

Что касается рассматриваемой задачи, то извлечение структурных элементов и их классификация даст дополнительное повышение качества в семантическом анализе текста.

Опора на модель реального мира

При анализе новостного текста удобно использовать модель отношений объектов предметной области. Состав объектов сохраняется для разных точек зрения, но структура и вид отношений между этими объектами отличаются. Именно эти отношения и определяют ту или иную точку зрения.

На лингвистическом уровне отношения можно выделить двумя способами:

- лексический – использование определенной лексики для оценки объектов;
- синтаксический – оценка объектов через специальное построение предложений.

Рассуждения на основе прецедентов

Как уже было сказано, основой предлагаемого решения является рассуждение на основе прецедентов (Case-based reasoning). Остановимся подробнее на истории и методологии данного подхода.

Работы Роджера Шэнка [3; 4] считаются истоками рассуждений на основе прецедентов. Шэнк предположил, что наше знание о мире организовано в виде пакетов памяти, хранящих эпизоды жизни. Такие пакеты (MOPs – memory organizations packets) и их элементы не изолированы, а пересекаются с нашими ожиданиями развития событий (сценариями). В свою очередь, MOPs образуют иерархию, где более общие пакеты объединяют более специфичные. Если MOP содержит ситуацию, в которой некоторая проблема была успешно решена, и человек находит себя в подобной ситуации, то он стремится использовать предыдущий опыт, чтобы найти решение. Таким образом, вместо следования общему набору правил, происходит повторное применение схемы решения в новых, но схожих условиях.

Существует четыре предположения о мире, которые составляют основу рассуждений на основе прецедентов.

1. Повторяемость: одинаковые действия, выполняемые в одинаковых условиях, приводят к одинаковым результатам.
2. Типичность: опыт имеет свойство самовоспроизводиться.
3. Согласованность: малые изменения в ситуации требуют малых изменений в интерпретации и решении.
4. Адаптивность: когда события повторяются, изменения склонны быть небольшими, а небольшие изменения легко сравнимы.

В общем случае процесс вывода на основе прецедентов состоит в следующих шагах: описание текущей проблемы, поиск максимально похожей решенной проблемы, получение решения для нее, адаптация решения к текущей проблеме, верификация нового решения, запоминание решения. Этот процесс называется «CBR-цикл» (рис. 1).

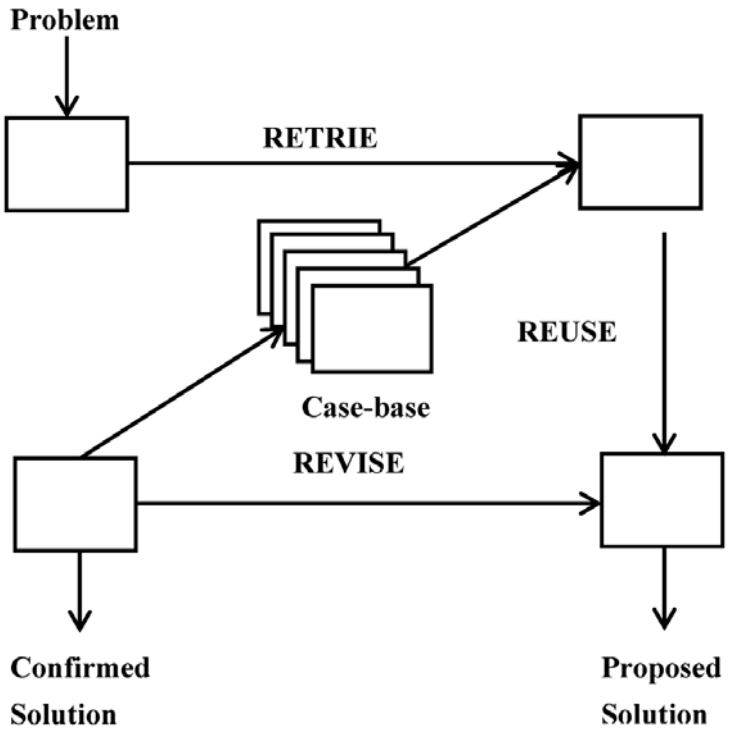

Рис. 1. Цикл CBR

Исследователи предлагают разные концепции и терминологию данного цикла. Согласно Колоднеру [5], этапы следующие.

- *Извлечение*: из библиотеки прецедентов извлекается наиболее близкий (подобный) прецедент для рассматриваемой проблемы.
- *Адаптация*: извлеченное решение адаптируется, чтобы лучше соответствовать новой проблеме.

- *Оценка решения*: адаптированное решение может быть оценено либо до его применения, либо после; в любом случае, если решение не подошло, то оно должно быть адаптировано еще раз, либо должны быть извлечены дополнительные решения.
- *Обновление базы прецедентов*: если решение прошло проверку успешно, новый прецедент добавляется в базу.

Аамодт и Плаза [6] предлагают схему «четырех RE»: Retrieve (извлечение), reuse (переиспользование), revise (пересмотр), retain (сохранение). Принципиально этот подход ничем не отличается от стандартной схемы.

В CBR выделяют 4 типа знаний: словарь, меру сходства, знание об адаптации и сами прецеденты. Словарь содержит информацию о свойствах, описывающих прецеденты. К типу знаний «мера сходства» относится как сама мера, так и такой способ организации прецедентов, чтобы можно было наиболее оптимально находить самый близкий прецедент к гипотетическим проблемам. Знание об адаптации содержит обычно правила модификации прецедентов и их оценки.

Более подробно остановимся на представлении прецедентов. В общем случае, прецедент состоит из следующих элементов:

- описание проблемы, т. е. состояние мира, когда возникает прецедент;
- решение проблемы и / или результата, т. е. состояние мира после обработки прецедента.

Прецеденты, которые сравнивают проблемы и их решения, могут быть использованы для адаптации решений к новым проблемам, а прецеденты, которые содержат описание проблемы и результат, могут быть использованы для оценки нового решения. Прецеденты могут быть представлены как векторы свойств или как любой формализм искусственного интеллекта, например фреймы, объекты, предикаты, семантические сети и т. п.

Отдельной проблемой является индексирование прецедентов в базе. Универсального подхода не существует. Есть несколько рекомендаций относительно того, какими свойствами должен обладать хороший индекс: предсказуемость, распознаваемость, абстрактность и отчетливость.

Важным аспектом проектирования CBR-систем является структура хранилища прецедентов. Существует три основных подхода к организации прецедентов.

Плоская организация – это простейший подход, представляющий собой плоский список прецедентов. Он легок в построении, обеспечивает быструю вставку новых прецедентов, но обладает плохими характеристиками производительности при поиске прецедентов. Для больших баз такой подход неприемлем.

Кластерная организация предполагает хранение в группах похожих друг на друга прецедентов. Поиск в такой структуре выполняется быстрее, чем в плоской организации, однако добавление и удаление требует реорганизации кластеров, что в некоторых случаях может быть очень дорогостоящей операцией.

Иерархическая организация представляет собой сетевую структуру категорий прецедентов. В лучшем случае такой подход предполагает наилучшую производительность при поиске прецедентов и достаточную при вставке / удалении. В худшем случае операции вставки и удаления могут привести к сложной реорганизации иерархии. Кроме того, необходимы дополнительные затраты на поддержку иерархии.

Далее опишем процедуры извлечения и адаптации прецедентов. Эти операции являются наиболее значимыми в рассуждениях на основе прецедентов.

Рассуждение на основе прецедентов во многом относится к рассуждению по аналогии, поэтому процедура извлечения прецедента может быть названа «вытаскиванием аналога». Самая простая форма извлечения – это поиск первого ближайшего соседа, при котором выполняется оценка близости ко всем прецедентам в базе и возвращается только один наиболее близкий. При большом объеме базы это может быть очень дорогостоящей операцией. Для решения этой проблемы прецедентам присваивается предвыборный приоритет. Это решает проблему сложности поиска, однако уменьшает точность извлечения.

Другой способ увеличения скорости извлечения – ранжирование прецедентов. Самое простое ранжирование достигается статистическими методами, т. е. вычислением вероятности для каждого прецедента быть извлеченным при поиске.

Помимо извлечения одного ближайшего соседа используется множественное извлечение. Это более сложная задача, связанная с определением оптимального количества извлекаемых прецедентов, однако потенциально результат ее выполнения может быть более эффективен, чем в случае с извлечением одного прецедента.

Что касается адаптации, существует два основных подхода.

1. Структурная адаптация. К решению непосредственно применяются правила адаптации. Если решение содержит единичное значение или коллекцию независимых значений, то структурная адаптация представляет собой изменение этих значений соответствующим образом.

2. Деривационная адаптация. Новое решение создается с помощью алгоритмов, методов или правил, которые сгенерировали оригинальное решение. Другими словами, происходит перегенерация прецедента.

Далее будет рассмотрен подход применения рассуждения на основе прецедентов к поставленной задаче. Однако перед тем, как сделать это, необходимо внести комментарии о форме представления текста новости.

Представление новости

Как говорилось выше, существует разные способы представления текстовой информации для компьютерной обработки. Один из самых популярных подходов, TF-IDF [7], заключается в нахождении слов, которые чаще всего встречаются в анализируемом тексте, но реже всего встречаются в коллекции документов, к которым принадлежит текст.

Метод «мешок слов» (Bag of words) представляет собой представление текста в виде вектора слов, встречающихся в тексте. Имея векторы нескольких текстов, можно производить над ними векторные операции, например вычислять косинусную меру, и таким образом определять «близость» текстов.

Развитие этого метода, технология word2vec, предлагается компанией Google [8]. В данном случае нейронная сеть, обучаясь на большом корпусе текстов, строит векторное пространство слов, на котором можно производить семантически значимые операции над полученными векторами. В данном случае речь идет не только о вычислении близости и / или коллинеарности векторов, как это происходит в случае Bag of words, но и, например, семантически значимые операции вычитания и сложения, типа «Король – мужчина + женщина = королева».

Перечисленные методы сильны своей статистической составляющей, простотой интерпретации результатов и удобной формой представления текста. Однако при этом сложно сказать, что они приемлемым образом отображают семантико-содержательную сторону текста. Для анализа новостей крайне важно получить ответ на то, какой субъект над каким объектом производит какие действия и в какой последовательности. В идеале желательно извлекать еще и предпосылки и причины действий, но оставим это за рамки данной статьи.

С учетом сказанного предлагается сценарно-ориентированный подход представления текста. Смысл состоит в том, чтобы разбить текст новости на множества предложений (возможно, состоящих и из одного предложения), каждое из которых является реализацией одного из заранее заданных сценариев. При этом в каждом сценарии можно выделить субъект, объект, предикат.

Рассуждение на основе прецедентов
как метод поиска аргументов и контраргументов

После представления текста в виде последовательности сценариев необходимо сделать важный шаг, а именно смоделировать рассуждение о тексте, причем постараться найти несколько точек зрения на рассматриваемое содержание. Для этого предлагается найти аргументы и контраргументы среди корпуса имеющихся новостей. Как основной механизм этой процедуры возьмем рассуждение на основе прецедентов. Опишем основные элементы рассматриваемого рассуждения на основе прецедентов. Для полного описания подхода необходимо описать следующее: словарь, меру сходства, адаптация, прецеденты.

Словарь содержит список всех возможных сценариев и список типов сценариев с точки зрения структуры новостного текста (см. структура новостного текста).

Прецедент имеет следующую структуру (см. таблицу):

- *проблема* – текст новости, представленный в виде последовательности сценариев;
- *решение* – вывод о новости в виде схемы «субъект – предикат – объект».

Структура прецедента

Проблема	Сценарий 1	Тип сценария	Субъект	Предикат	Объект
	Сценарий 2	Тип сценария	Субъект	Предикат	Объект
	…	…	…	…	…
	Сценарий *N*	Тип сценария	Субъект	Предикат	Объект
Решение	Субъект – Предикат – Объект				

Структура хранения прецедентов – плоская. Индексация производится по сценарию, соответствующему главному событию новости.

Мера сходства определяется как размер наибольшей общей подпоследовательности (НОП) последовательностей сценариев двух прецедентов. Сценарии считаются совпадающими, если они имеют один и тот же тип, а также совпадают объект или субъект сценариев. При этом если размеры НОП совпадают с несколькими прецедентами, то более близкий прецедент определяется с позиции лексикографического порядка (ввиду принципа «перевернутой пирамиды»). Другими словами, совпадение более ранних сценариев ценнее, чем более поздних. Извлечение происходит методом *k* ближайших соседей.

Адаптация заключается в синтезе выводов найденных прецедентов со сценариями, которые не совпадают. В идеале необходимо найти *аргументы* – прецеденты, дополняющие рассматриваемый прецедент, и *контраргументы* – опровергающие его. Далее процедуру можно повторять рекурсивно, получая таким образом цепочки рассуждений, каждая из которых является отдельной интерпретацией новости (рис. 2).

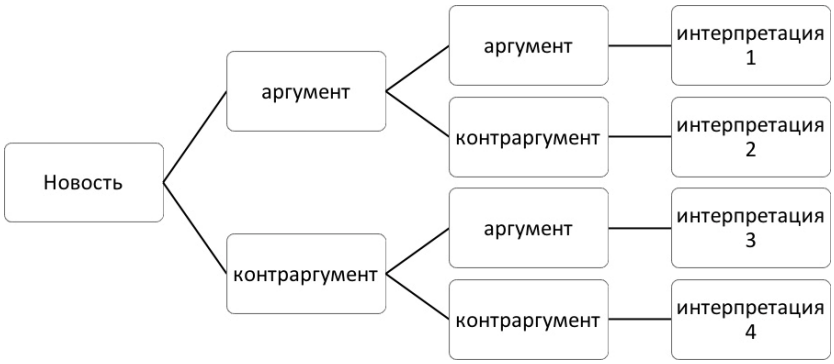

Рис. 2. Цепочки рассуждений

Заключение

В статье рассмотрен подход к моделированию рассуждений о новостном тексте, который служит основой для решения более общей задачи моделирования распространения новостных сообщений в Интернете. Рассмотрены методологические основы, на которых базируется подход, а именно рассуждения на основе прецедентов и сценарно-ориентированный подход к анализу текстов на естественном языке. Особенностью предлагаемого подхода является порождение разных интерпретаций рассуждений, что в будущем позволит более гибко решать задачу моделирования распространения сообщений.

Список литературы

1. *Цыбикова Н. С.* Диктемная структура текста интернет-новостей // Вестн. ВятГГУ. 2011. № 3 (39). С. 76–79.
2. *Ягунова Е. В., Пивоварова Л. М.* Исследование структуры новостного текста как последовательности связных сегментов // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог». 2011. № 10 (17). С. 273–287.
3. *Schank R. C.* Dynamic memory: A theory of reminding and learning in computers and people. Cambridge: Cambridge University Press, 1982.
4. *Schank R. C.* Memory-based expert systems. Technical Report (# AFOSR. TR. 84-0814). New Haven: Yale University, 1984.
5. *Kolodner J. L.* An introduction to case-based reasoning // Artificial Intelligence Review. 1992. № 1 (6). P. 3–34.
6. *Aamodt A., Plaza E.* CBR: foundational issues, methodological variations and system approaches // AI Communications. 1994. № 1 (7). P. 39–59.
7. *Jones K. S.* A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation. 2004. № 5 (60). P. 493–502.
8. *Mikolov T. et al.* Distributed Representations of Words and Phrases and their Compositionality // Nips. 2013. P. 1–9.

Материал поступил в редколлегию 20.04.2017

P. V. Myznikov

Novosibirsk State University
4 Lyapunov Str., Novosibirsk, 630090, Russian Federation

miznikov72@gmail.com

CASE-BASED REASONING IN AUTOMATIC ANALYSIS OF NEWS TEXTS

Case-based reasoning is considered in the tasks of analysis and modelling of messages distribution in Internet. The approach proposed allows create sequences of reasoning about news content. Special attention is paid to searching arguments and counterarguments in similar texts for construction of several viewpoints on the news.

Keywords: reasoning modelling, cases, natural languages processing, text mining.

References

1. Tsybikova N.S. Dictemic structure of Internet newstext // Vyatka State University Scientific Journal. 2011, vol. 39, iss. 3, p. 76–79.
2. Yagunova E.V., Pivovarova L.M. A study of the news text structure as a consequnce of connected segments // Proceedings of the Annual International Conference “Dialogue” (2011). Moscow, 2011. p. 273–287.
3. Aamodt A., Plaza E. CBR: foundational issues, methodological variations and system approaches // AI Communications. 1994. № 1 (7). p. 39–59.
4. Jones K.S. A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation. 2004. № 5 (60). p. 493–502.
5. Kolodner J.L. An introduction to case-based reasoning // Artificial Intelligence Review. 1992. № 1 (6). p. 3–34.
6. Mikolov T. and others. Distributed Representations of Words and Phrases and their Compositionality // Nips. 2013. p. 1–9.
7. Schank R.C. Dynamic memory: A theory of reminding and learning in computers and people / R.C. Schank, Cambridge: Cambridge University Press, 1982.
8. Schank R.C. Memory-based expert systems. Technical Report (# AFOSR. TR. 84- 0814) / R.C. Schank, New Haven: Yale University, 1984.