# ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ ИЗ НОВОСТНЫХ ТЕКСТОВ С ПОМОЩЬЮ МОДЕЛИРОВАНИЯ РАССУЖДЕНИЙ НА ОСНОВЕ ПРЕЦЕДЕНТОВ

П.В. Мызников
Научный руководитель – Д.Е. Пальчунов
Новосибирский Государственный Университет
p.myznikov@g.nsu.com

**Введение**

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

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

1) оценка охвата аудитории новостного сообщения,

2) оценка вероятности попадания сообщения в определённый новостной источник,

3) оценка степени интереса пользователей к определённой 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 задачах, а именно: интерпретация текста с разных точек зрения. Разные люди по-разному воспринимают информацию: учёт этой особенности позволит более точно моделировать распространение сообщений в Интернете. Этой проблеме посвящено данное исследование.

**Сценарно-ориентированный подход обработки текста**

Важным моментом является метод формализации новостного текста. Существует несколько способов отображения текста на естественном языке для компьютерной обработки. К самым распространённым можно отнести TF-IDF, Bag of words, word2vec.

Перечисленные методы сильны своей статистической составляющей, простотой интерпретации результатов и удобной формой представления текста. Однако, при всём при этом, сложно сказать, что они должным 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. Кроме этого, необходимо учитывать особенности новостного текста с точки зрения лингвистики. В идеале желательно извлекать ещё и предпосылки и причины действий, но оставим это за рамки данной статьи.

Ввиду этого 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.

**Рассуждения на основе прецедентов**

Основой предлагаемого решения является рассуждения на основе прецедентов (case-based reasoning). Описание этого подхода содержится в работах Р. Шэнка [1][2] и Дж. Колоднера [3]. Суть подхода состоит в решении новых проблем путём адаптации решений похожих проблем в прошлом. Процесс вывода на основе прецедентов состоит и четырёх шагов и называется CBR-циклом:

- извлечение: из библиотеки прецедентов извлекается наиболее близкий (подобный) прецедент для рассматриваемой проблемы;
- адаптация: извлечённое 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 извлечены дополнительные решения;
- обновление базы прецедентов: если решение прошло проверку успешно, новый прецедент добавляется в базу.

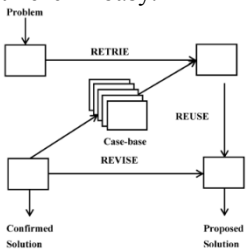


Рис. 2. CBR-цикл

**Схема моделирования рассуждения**

С точки зрения рассуждений на основе прецедентов, необходимо описать следующие элементы: словарь, меру сходства, процедуру адаптации и форму прецедентов.

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

Прецедент имеет следующую структуру:

- проблема: текст новости, представленный в виде последовательности сценариев;
- решение: вывод о 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. Другими словами, совпадение более ранних сценариев ценнее, чем более поздних. Извлечение происходит методом k ближайших соседей.

Важнейшим моментом в подходе является использование онтологических моделей [4][5]. На их основе строится процедура адаптации. Согласно соответствующей онтологии, из кластера, которому принадлежит анализируемый прецедент, производится поиск аргументов и контраргументов. Задавая разные онтологические модели, возможно моделировать рассуждения разных точек зрения, а соответственно, и разные интерпретации одного и того же текста.

Теоретико-модельный подход к описанию прецедентов был разработан и применен к решению задач информационной безопасности [6], а также в медицине [7][8]. Этот подход основан на применении булевозначных и нечетких моделей [9][10]. В [9] введено понятие прецедентной модели и показано, что она является булевозначной моделью.

Обобщая выше сказанное, получаем следующий CBR-цикл:

- анализируемый текст представляется в виде последовательности реализаций сценариев,
- извлекается текст, наиболее близкий анализируемому с точки зрения сценарной структуры и содержания,
- адаптация рассуждения о найденном тексте к рассуждению об обрабатываемом тексте производится на основе заранее заданной онтологической модели соответствующей предметной области,
- сформулированное рассуждение сохраняется в базе прецедентов

**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. Особенностью предлагаемого подхода является порождение разных интерпретаций рассуждений, что в будущем позволит более гибко решать задачу моделирования распространения сообщений. Указана перспективность использования онтологического метода для внесения в моделирование рассуждений специфику контекста сообщения.

**Список используемых источников**

1. Schank R.C. Dynamic memory: A theory of reminding and learning in computers and people / R.C. Schank, Cambridge: Cambridge University Press, 1982.
2. Schank R.C. Memory-based expert systems. Technical Report (# AFOSR. TR. 84- 0814) / R.C. Schank, New Haven: Yale University, 1984.
3. Kolodner J.L. An introduction to case-based reasoning // Artificial Intelligence Review. 1992. № 1 (6). С. 3–34.
4. Пальчунов Д.Е. Решение задачи поиска информации на основе онтологии. Бизнес-информатика. 2008. № 1 (3). С. 3-13.
5. Пальчунов Д.Е. Моделирование мышления и формализация рефлексии. Ч.2. Онтологии и формализация понятий. Философия науки. 2008. № 2 (37). С. 62-99.
6. Yakhyaeva G.E., Yasinskya O.V. Application of Case-based Methodology for Early Diagnosis of Computer Attacks // Journal of Computing and Information Technology - CIT 22, 2014, 3, 145–150.
7. Найданов Ч.А., Пальчунов Д.Е., Сазонова П.А. Теоретико-модельные методы интеграции знаний, извлеченных из медицинских документов. Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2015. Т. 13. № 3. С. 29-41.
8. Пальчунов Д.Е., Яхъяева Г.Э., Ясинская О.В. Применение теоретико-модельных методов и онтологического моделирования для автоматизации диагностирования заболеваний. Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2015. Т. 13. № 3. С. 42-51.
9. Пальчунов Д.Е., Яхъяева Г.Э. Нечеткие алгебраические системы. Сибирский журнал чистой и прикладной математики. 2010. Т. 10. № 3. С. 76-93.
10. Пальчунов Д.Е., Яхъяева Г.Э. Нечёткие логики и теория нечётких моделей. Алгебра и логика, 54, № 1, 2015, с. 109-118.