Извлечение терминов из русскоязычных текстов при помощи графовых моделей

Усталов Дмитрий Алексеевич

dmitry@eveel.ru

УрФУ, Екатеринбург, Россия

Аннотация. Статья посвящена вопросу извлечения терминов из текстов на русском языке при помощи графовых моделей. Описан и экспериментально исследован алгоритм решения данной задачи. Сформулированы требования и рекомендации к применению алгоритма в задачах обработки русского языка.

Ключевые слова: извлечение терминов; извлечение ключевых слов; анализ графов; обработка естественного языка; компьютерная лингвистика; корпусная лингвистика.

Введение

В современном Интернете наблюдается тенденция к постепенному внедрению структурированных данных в Web-страницы посредством микроформатов, семантических сетей, особых элементов разметки гипертекста. Это даёт возможность использовать ресурсы Интернета не только как корпус текстов, но и как полноценную базу данных.

При всех преимуществах стека технологий Semantic Web, в русскоязычном сегменте Сети существует проблема, которая несколько лет назад наблюдалась и в англоязычном Интернете — порочный

круг «недостаточно данных → их обработка бесполезна → публиковать данные нецелесообразно».

Одно из эффективных решений этой проблемы заключается в автоматическом структурировании существующих данных: из неструктурированного текста могут быть выделены теги, именованные сущности, семантические отношения, и т. д. Имеется коммерческая ценность таких мероприятий в виде дополнительной информации для принятия решений, особенно при продвижении бизнеса в социальных СМИ и поисковых системах.

Под *термином* в данной работе понимается слово или словосочетание, способное в совокупности с другими терминами представлять текст.

Задача выделения терминов из текста возникает в библиотечном деле, лексикографии и терминоведении, а также в информационном поиске. Выделенные автоматическим образом слова и словосочетания могут использоваться для создания и развития терминологических ресурсов, а также для эффективной обработки документов: индексирования, реферирования, классификации [1].

Графовые модели представляют большой интерес для области обработки естественного языка благодаря своей универсальности и эффективности основанных на них алгоритмов, что продемонстрировано в работах [2, 3, 4].

Большой проблемой обработки русского языка является недоступность необходимых словарей, корпусов, тезаурусов и программного обеспечения, что предполагает создание и применение инструментов, функционирующих на основе подхода «чистой доски».

Несмотря на независимость графовых моделей от языка [4], целесообразно уточнить параметры модели и её поведение, поскольку грамматика русского языка заметно отличается от грамматики английского языка, хорошо изученного в данной области.

Целью данной работы является уточнение параметров и синтез рекомендаций при использовании графовых моделей для обработки русского языка путём решения задачи извлечения терминов.

1 Алгоритм

Для решения исходной задачи необходимо построить граф на основе текста, выполнить ранжирование его вершин, а также провести сборку словосочетаний на основе вершин, имеющих наибольший вес.

Построение графа. Будет выполняться построение взвешенного неориентированного графа в виде $G = (V, E)$, где V — множество слов, E — множество связей между ними.

В качестве V можно принять множество всех уникальных лемм исходного текста. Поскольку большинство терминов являются именными группами [1], то множество слов стоит ограничить только леммами, образованными от имён существительных и прилагательных.

Множество E строится путём последовательного сканирования текста заданным окном из $N \in [2, 10]$ слов. На каждой итерации для пары слов вычисляется величина связи $WC(w_1, w_2)$, обратно зависящая от расстояния между словами:

$$WC(w_1, w_2) = \begin{cases} 1 - \frac{d(w_1,w_2)-1}{N-1}, & \text{если } d(w_1, w_2) \in (0, N), \\ 0, & \text{если } d(w_1, w_2) \geqslant N, \end{cases}$$

где w_1 и w_2 — слова, $d(w_1, w_2)$ — расстояние между словами, N — размер окна.

Слова, для которых величина $WC(w_1, w_2)$ приняла нулевое значение, не включаются во множество вершин графа. Для определения расстояния между двумя словами достаточно воспользоваться позиционной мерой:

$$d(w_1, w_2) = \big|p(w_1) - p(w_2)\big|, \text{ при } w_1 \neq w_2,$$

где w_1 и w_2 — слова, $p(w)$ — порядковый номер слова w в тексте.

Основанием для вычисления величины $WC(w_1, w_2)$ служит наблюдение [2], что между двумя рядом стоящими словами часто существует семантическое отношение. Это необходимо для обеспечения связности представления текста в виде графа. Чем выше расстояние $d(w_1, w_2)$, тем ниже вероятность существования такого отношения.

При обработке графа работа ведётся исключительно с одиночными именами существительными и прилагательными. Объединение этих слов в словосочетания будет выполнено на этапе сборки словосочетаний.

Ранжирование вершин графа. После генерации графа G необходимо вычислить TextRank — значение стационарного распределения случайного блуждания для каждой вершины $t \in V$ с учётом весов связей [2]:

$$TR(t_i) = (1 - d) + d \cdot \sum_{t_j \in In(t_i)} \frac{w_{ji}}{\sum_{t_k \in Out(t_j)} w_{jk}} \cdot TR(t_j),$$

где d — фактор затухания, $In(t)$ — множество вершин, входящих в t, $Out(t)$ — множество вершин, исходящих из t, w_{ij} — вес ребра (t_i, t_j). Для неориентированного графа принято $In(t) \equiv Out(t)$.

После вычисления TextRank необходимо составить множество C кандидатов в термины из первых T слов из списка вершин, упорядоченных по убыванию значения TextRank. Существует ряд подходов к определению $T \equiv |C|$: принять постоянное значение T, использовать $T = \frac{1}{3}|V|$, и т. д.

Сборка словосочетаний. Необходимо извлечь из текста все последовательности слов, состоящих из элементов множества C. Поскольку $\forall t \in C : \exists! TR(t) \in \mathbb{R}$, то общий вес извлечённой последовательности определяется суммой весов всех составляющих её слов.

При сборке словосочетаний стоит учитывать два момента: 1) последовательность обязана иметь в составе хотя бы одно имя существительное; 2) при обнаружении вложенности одной последовательности в другую рассматривается только последовательность с бóльшим весом.

Выделенные последовательности из одного и более слов можно рассматривать в качестве терминов исходного текста.

Пример. На рис. 1 приведён граф текста аннотации данной статьи. При выделении слов использовалось окно размером в пять слов.

Рис. 1. Граф текста аннотации данной статьи

Извлечено восемь слов слов из упорядоченного списка ранжированных вершин графа: *задача*^{0,094}, *русский*^{0,084}, *алгоритм*^{0,083}, *язык*^{0,076}, *извлечение*^{0,064}, *обработка*^{0,063}, *термин*^{0,062} и *вопрос*^{0,060}.

Из этих слов автоматически собрано три термина: *задача обработки русского языка*^{0,317}, *вопрос извлечения терминов*^{0,186} и *ал-*

горитм^{0,083}. Данные термины вполне соответствуют тексту аннотации.

2 Эксперимент

Для более детального исследования алгоритма выполнена обработка материалов проекта «Открытый корпус» [5], морфологически размеченных программой TreeTagger[1]:

— пять коротких текстов (1–2 абзаца);
— пять текстов среднего объёма (2–7 абзацев);
— пять крупных текстов (7–15 абзацев).

Эксперимент выполнялся путём экспертной бинарной оценки каждого извлечённого термина на предмет соответствия тематике содержащего его документа. Предварительного извлечения терминов вручную не проводилось, что не позволяет количественно определить полноту результата.

Результаты эксперимента приведены в табл. 1, где приняты следующие обозначения: N — выбранное окно, P — точность, σ — стандартное отклонение значения точности, W — среднее количество слов в выделенном термине. Индексы S, M, L принадлежат коллекциям коротких, средних, крупных текстов, соответственно.

Значения $N > 4$ не рассматривались, поскольку известно увеличение приводит к заметной деградации точности извлечения терминов [2]. При обработке коллекции документов принято $T = \frac{1}{3}|V|$.

Табл. 1. Результаты сравнения на коллекциях текстов

N	P_S	σ_S	W_S	P_M	σ_M	W_M	P_L	σ_L	W_L
2	50%	15%	4,4	44%	19%	9,6	47%	7%	7,6
3	59%	22%	5,8	37%	12%	13,4	50%	13%	12,0
4	55%	27%	5,8	38%	16%	13,8	49%	10%	13,8

Анализ. Обработка трёх коллекций документов показала заметный разброс точности, поэтому анализ проведён с учётом пессимистичного сценария ($P - \sigma$) среди лучшего результата для данной коллекции.

Коллекции документов показали лучший результат 35% для коротких, 25% для средних, 40% для крупных текстов при $N = 2$.

Результат обработки коротких и крупных текстов превышает точность ≈31%, показанную в работе [2]. Важно отметить, что более

[1] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

развитые алгоритмы на основе графовых моделей способны обеспечить точность ≈75% [4].

Из-за разницы в методиках тестирования, подтверждение полученного результата требует отдельного статистического теста на более крупной коллекции документов. Можно воспользоваться методикой оценки, аналогичной [1], где по нечёткой формальной оценке «среднего списка» наблюдается точность ≈22%.

Изучение полученных списков терминов выявило две проблемы, заметно снижающие точность результата: недостатки морфологической разметки и сложности при сборке словосочетаний.

Морфологическая разметка. Некорректная или отсутствующая морфологическая разметка некоторых слов приводит либо к появлению лишних вершин графа, либо к отсутствию необходимых.

Например, анализатор TreeTagger некорректно обрабатывает слова, содержащие букву «ё», определяя такие слова как имена существительные. Также наблюдаются проблемы с определением имён собственных, из-за чего в граф не попадает много потенциально важных вершин.

Проблема может быть решена путём усовершенствования средств предварительной обработки текста или частично решена заменой всех вхождений буквы «ё» на «е».

Сборка словосочетаний. Недостаточность ограничений на этапе сборки словосочетаний приводит к появлению заметного количества бессмысленных строк, собранных из вершин с большим весом.

Например, в результатах работы наблюдались такие словочетания как «год рабочего» вместо «год рабочего стажа», и «г углекислого газа» вместо «углекислый газ».

Основное решение проблемы состоит в учёте вложенности и частоты встречаемости терминов в тексте. Здесь возникает проблема: при обнаружении словосочетаний «пшеничный хлеб» и «сладкий хлеб» стоит выделить общий термин «хлеб», но в ситуации «Иван Иванов» и «Иван Петров» целесообразно оставить оба варианта.

Вероятно, имеет смысл рассмотреть возможность привлечения дополнительных словарей и различных эвристических алгоритмов.

Заключение

Исследование описанного алгоритма извлечения терминов показало заметный разброс точности и выявило проблемы, возникающие на этапах морфологической разметки и сборки словосочетаний.

Экспериментально определён оптимальный размер окна при составлении графа текста — два слова, как для английского языка [2].

Учёт этих рекомендаций и повторный статистический тест на более крупной коллекции документов позволит приблизиться к результатам, продемонстрированным в работе [4].

Благодарности

Автор благодарит всех участников проекта «Открытый корпус»[2] за проделанную работу по сбору и разметке свободных текстов на русском языке. При подготовке статьи использовалось Gephi[3] — свободное программное обеспечение для анализа графов.

Список источников

1. Браславский, П., Соколов, Е. Сравнение пяти методов извлечения терминов произвольной длины. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». — Вып. 7 (14). — М.: РГГУ, 2008. — С. 67–74.
2. Mihalcea, R., Tarau, P. TextRank: Bringing Order into Texts. // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. — 2004. — Vol. 4. — № 4. — P. 404–411.
3. Grineva, M., Grinev, M., Lizorkin, D. Extracting Key Terms From Noisy and Multi-theme Documents. // Proceedings of the 18th International Conference on World Wide Web. — 2009. — P. 661–670.
4. Litvak, M., Last, M., Kandel, A. DegExt: a language-independent keyphrase extractor. // Journal of Ambient Intelligence and Humanized Computing. / Springer. — 2012. — P. 1-11.
5. Бочаров, В., Грановский, Д. Программное обеспечение для коллективной работы над морфологической разметкой корпуса. // Труды международной конференции «Корпусная лингвистика – 2011». 27–29 июня 2011 г., Санкт-Петербург. — СПб.: С.-Петербургский гос. университет, Филологический факультет, 2011. — 348 с.

[2] `http://opencorpora.org`
[3] `http://gephi.org`