ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ СЛОВ: ГРАФООРИЕНТИРОВАННЫЙ ПОДХОД

Мазура И. А., Гуринович А. Б.
Кафедра информационных технологий автоматизированных систем,
Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: irynamazura22@gmail.com, gurinovich@bsuir.by

Задача автоматического выделения ключевых слов из текста возникает во многих областях. Актуально, как исследование существующих алгоритмов, так и анализ методов, позволяющих их оптимизировать.

Введение

Извлечение ключевых слов является важной задачей, лежащей на стыке таких областей знаний, как интеллектуальный анализ текста (Text Mining), информационный поиск (Information Retrieval) и обработка естественного языка (Natural Language Processing). Под ключевыми словами понимаются слова или фразы, которые наилучшим образом характеризует содержание исследуемого текста. Задача автоматического определения ключевых слов представляет собой необходимый этап обработки текста для решения таких задач, как: создание и развитие терминологических ресурсов, автоматический информационный поиск, аннотирование, классификация и кластеризация документов, суммаризация текста и другие.

I. Краткий обзор подходов к извлечению ключевых слов

Анализ литературы в области извлечения ключевых слов выявил большое число методов и их модификаций, однако общепринятой в научном сообществе классификации данных подходов на текущий момент не существует. Действительно, решение задачи автоматического выделения ключевых слов ведется одновременно по двум направлениям. С одной стороны, методы различаются по типу математического аппарата распознавания ключевых слов (статистические, методы на основе машинного обучения, структурные), с другой — по типу используемых (или не используемых) лингвистических ресурсов (словари, онтологии и тезаурусы, корпуса текстов). Наиболее простым статистическим методом извлечения ключевых слов является метод ранжирования всех словоформ по частоте. При подсчете частоты употребления ключевого слова учитываются все его словоформы в тексте. Создаваемые на основе данного подхода алгоритмы являются недостаточно точными, т.к. признак частотности ключевых слов не является превалирующим [1].

Для повышения корректности автоматического извлечения ключевых слов, статистический метод дополняется лингвистическими процедурами (морфологическим, синтаксическим или семантическим анализом). Такие методы могут требовать или не требовать корпусов текстов. Использование корпуса текстов получило достаточно широкое распространение, однако отсутствие таких корпусов для каждой конкретной предметной области в реальной жизни делает применение таких корпусных моделей и методов весьма проблематичным. Методы на основе машинного обучения рассматривают задачу извлечения ключевых слов как задачу классификации — вычисление вероятности отнесения слова к ключевому на основе обучающей выборки — корпуса документов с размеченными ключевыми словами. В основе структурных методов лежит представление о тексте как системе семантически и грамматически взаимосвязанных элементов-слов, которые, в свою очередь, характеризуются набором лингвистических признаков. Здесь могут быть выделены два подкласса — графовые и синтаксические (шаблонные) методы.

II. Графоориентированный подход

Выбор того или иного алгоритма извлечения ключевых слов обуславливается в первую очередь естественным языком, спецификой исследуемой темы, объемом анализируемого текста. Графовые модели представляют большой интерес для области обработки естественного языка благодаря своей универсальности (не зависят от естественного языка) и эффективности основанных на них алгоритмов. Графовые методы не предполагают использование лингвистических ресурсов для настройки критериев принятия решений при распознавании ключевых слов. Вместо этого, в работе алгоритмов подразумевается контекстно-независимое выделение ключевых слов, что является оптимальным решением для гомогенных по функциональному стилю корпусов текстов, например, научных работ или нормативных актов, а также работ, посвященных новым, развивающимся темам, для которых не существует разработанных словарей, тезаурусов и т.п.

Итак, в основе графовых моделей лежит процедура построения графа, в вершинах кото-

рого стоят лексические единицы (слова или предложения), а отношения между ними представлены в виде ребер графа. Отношение между лексическими единицами может быть основано на различных принципах, наиболее распространенными из которых являются:

- Отношение совместной встречаемости — связанные слова встречаются в тексте внутри окна фиксированного размера; связаны все слова внутри предложения, параграфа или документа.
- Семантическое отношение — связанные слова имеют одинаковое значение, синонимы, антонимы, омонимы и др.

Для вершин полученного графа вычисляется мера центральности как индикатор определения наиболее значимых вершин внутри графа. Центральность вершины v — это мера, отражающая то, насколько вершина v участвует в процессе распространения информации между остальными вершинами в графе. В области извлечения ключевых слов используются различные меры центральности, на основе которых производится ранжирование слов текста [3].

III. Алгоритм TextRank

В основе TextRank лежит процедура построения взвешенного графа, в вершинах которого стоят лексические единицы (слова или предложения), а ребра взвешены в соответствии с силой связи между ними. После того, как произведена предобработка текста, производится построение взвешенного неориентированного графа $G = (V, E)$, где V — множество уникальных лексических единиц (вершины); E — множество связей между ними (ребра) [2].

В качестве меры связи между словами TextRank использует отношение совместной встречаемости: две вершины соединяются ребром, если их лексические единицы встречаются вместе внутри окна из N слов, $N \in [2, 10]$. Формула для определения веса каждого ребра:

$$w_{ij} = \begin{cases} 1 - \dfrac{d(w_i, w_j) - 1}{(N-1)}, & \text{если } d(w_i, w_j) \epsilon (0, N) \\ 0, & \text{если } d(w_i, w_j) \geq (0, N) \end{cases}$$

На следующем этапе по итеративной формуле вычисляется TextRank (TR), получаемый случайным блужданием для каждой вершины из V:

$$TR(V_i) = (1-d) + d \cdot \sum_{V_j \epsilon In(V_i)} \frac{w_{ij}}{\sum_{V_k \epsilon Out(V_j)} w_{jk}} TR(V_j),$$

Начальное значение TR для каждой из вершин предполагается равным 1. Алгоритм повторяется до достижения порогового уровня сходимости.

В соответствии с итоговыми значениями TR вершины графа ранжируются, отбираются T «лучших» вершин (с наибольшим значением TR). Ключевые фразы получаются путем извлечения из текста последовательностей, состоящих из T-лучших слов [4].

IV. Заключение

Алгоритм TextRank показывает адекватные результаты на обработке текстов сравнительно большого размера. В работе алгоритма были выявлены определенные сложности в сборке словосочетаний: «склеивание» слов в фразы происходит в полуручном режиме (необходимо участие эксперта), так как автоматическая сборка, во-первых, способна вывести только наборы словлемм, во-вторых, недостаточность ограничений на этапе сборки словосочетаний приводит к появлению заметного количества бессмысленных строк, собранных из вершин с большим весом. Основное решение проблемы состоит в учете вложенности и частоты встречаемости терминов в тексте.

Несмотря на большое число методов извлечения ключевых слов, до настоящего времени не разработана последовательная методика обнаружения ключевых слов человеком. Экспериментально подтверждено, что эта операция выполняется людьми интуитивно и является личностно обусловленной. Так, в случае извлечения ключевых слов из небольшого числа документов, результаты будут чувствительны к количеству слов, которое – было принято решение – отсечь от общего числа слов-кандидатов в ключевые слова. Употребляя одни и те же слова в разных словосочетаниях и контекстах, авторы текстов уходят от четких границ терминов. Тем самым размывается их содержание, исчезает терминологический статус.

Актуальным на данный момент является вопрос о составлении объективной методики извлечения ключевых слов.

V. Список литературы

1. Beliga, S., Martincic-Ipsic, S., and Meštrović, A. An Overview of Graph-Based Keyword Extraction Methods and Approaches. Journal of Information and Organizational Sciences, 2015,39(1).
2. Page, S., Brin, S., Motwani, R, and Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford: Stanford University, 1998.
3. Automatic Keyword Extraction from Individual Documents [Электронный ресурс]. – Режим доступа: https://www.researchgate.net/publication/227988510 _Automatic_Keyword_Extraction_-d-from_Individual_Documents.
4. Understand TextRank for Keyword Extraction by Python [Электронный ресурс]. – Режим доступа: https://towardsdatascience.com/textrank-for-keyword-extraction-by-python-c0bae21bcec0.