УДК 004.912

СИСТЕМА ПОИСКА ДОКУМЕНТОВ, РЕЛЕВАНТНЫХ ЗАДАННОМУ ТЕКСТУ

Полушин Глеб Валерьевич

Национальный исследовательский университет «Высшая школа экономики», 614070, Россия, г. Пермь, ул. Студенческая, 38, polushin.gleb@mail.ru

Статья посвящена вопросам автоматизации процесса поиска документов релевантных заданному тексту. Поиск актуальной литературы является нетривиальной задачей, так как требует временных затрат и предъявляет определенные требования к читателю. В данной работе предложены методы, позволяющие минимизировать трудозатраты при поиске релевантной литературыи автоматизировать весь процесс, произведен анализ существующих программных решений по извлечению ключевых слов из текстов на русском языке. Описан способ извлечения ключевых фраз из текста на русском языке с использованием метрики TF-IDF. Также описан способ автоматизированного поиска актуальных документов по извлеченным ключевым словам с использованием бесплатной поисковой системы по полным текстам научных публикаций всех форматов и дисциплин GoogleScholar. На практике система может быть использована студентами, преподавателями и всеми, кто занимается научной деятельностью.

Ключевые слова: извлечение ключевых слов, автоматизированный поиск, обработка текста.

Введение

Одним из ключевых навыков в научной деятельности является поиск литературы по интересующей теме. Условно процесс поиска релевантной литературы можно разделить на два этапа: извлечение ключевых слов из текста и поиск литературы по ключевым словам. Каждый из этапов является самостоятельной задачей.

Процесс составления ключевых фраз текста очень трудоемок и занимает большое количество времени. Решением данной проблемы является автоматическое извлечение ключевых фраз из текста. Существуют различные алгоритмы и системы по извлечению ключевых фраз. Однако обычно такие системы не универсальны и могут работать с одним – двумя языками, так как большую роль в данном процессе играют особенности языка. Русский язык достаточно сложен для автоматической обработки, соответственно существует не так много систем по извлечению ключевых фраз из текстов на русском языке.

При поиске литературы по ключевым словампользователь обычно работает с электронной библиотекой, выполняя запрос по извлеченным ключевым словам. Существуют различные сайты, приложения, которые работают с известными базами данных, либо имеют свою базу данных научной литературы.

Во время изучения степени разработанности проблемы, было обнаружено отсутствие систем, которые бы автоматически выполняли оба этапа поиска литературы. Автоматизированы только отдельные этапы процесса, а не сам процесс, соответственно разработка системы поиска релевантной литературы является в настоящее время актуальной темой.

Анализ систем извлечения ключевых слов из текста

Системпо извлечению ключевых слов на русском языке достаточно мало. Рассмотрим некоторые их них.

ContentAnalyzer [1] – инструмент для анализа содержания тематических web-страниц в реальном времени, выделения списков ключевых слов и словосочетаний, построенияавтореферата текста документа. Функционирование ContentAnalyzer обеспечивается за счёт вычисления частоты термина или словосочетания в документе. Система анализирует текст в реальном времени, что очень удобно, однако методы выделения ключевых слов являются достаточно примитивнымии неэффективными.

TextAnalyst [2] – инструмент для поиска информации и анализа содержания текстов, имеющий возможность выделения ключевых слов. Функционирование TextAnalyst основано на применении методов обработки естественного языка в сочетании с методами машинного обучения. В процессе своей работы, система создает иерархическую нейронную сеть. Сетьсодержит несколько слоев: фрагменты, которые встречаются в тексте более одного раза, хранятся в тех нейронах, которые принадлежат к более высоким уровням сети.

Tesuçk [3]– это веб-сервис автоматического выделения ключевых слов и словосочетаний из текста на естественном языке. Также сервис позволяет произвести автореферирование текста. Сервис строит графовую модель текста, используя метод TextRank, который используется для вычисления весов вершин графа.

Описание предлагаемого подхода

Предложено разработать систему поиска документов, релевантных содержанию заданного текста. Система должна совмещать в себе два этапа поиска релевантной литературы: извлечение ключевых слов и поиск по ним литературы в базах данных научных публикаций. На вход системе подается статья, после обработки которой пользователю выводятся ключевые слова и фразы с предложением выбрать нужные для поиска в Интернете. После поиска пользователю выдаются ссылки на статьи по данной теме.

При извлечении ключевых фраз предложено выбрать все N-граммы длины 1 и 2. В случае если это одно слово, то это должно быть существительное. В случае двух слов

возможны два варианта: это может быть прилагательное и существительное или два существительных, причем второе в родительном падеже. Необходимо приводить термины в нормальную форму, чтобы избежать дублирования при ранжировании. Для ранжирования кандидатов предложено выбрать метрику TF-IDF (1), которая является самой распространенной мерой для расчета информативности терминов [4]. Вес термина пропорционален количеству употреблений данного термина в документе, и обратно пропорционален частоте употребления в других документах коллекции.

$$TF - IDF = TF * IDF \quad (1)$$

TF (Term Frequency) – частота термина в анализируемом документе, отношение числа вхождения термина к количеству терминов в документе (2):

$$TF = \frac{n_i}{\sum n_k} \quad (2)$$

IDF (Inverse Document Frequency) – инвертированная частота документа, частота с которой термин встречается в других документах коллекции (3):

$$IDF = \log \frac{N}{df}, \quad (3)$$

где N – общее количество документов в коллекции (корпусе), df – количество документов, содержащих термин. Выбор основания логарифма не имеет значения, так как не влияет на соотношение весов терминов. Для расчета IDFнеобходимо произвести предварительную обработку корпуса научных текстов на русском языке. В процессе поддержки системы нужно периодически производить перерасчет частот терминов, для поддержания актуальности.

Для непосредственного поиска документов по ключевым словам предложено использовать инструменты Академии Гугл (GoogleScholar) [5], так как ее индекс включает данные из большинства рецензируемых онлайн журналов крупнейших научных издательств. Для получения результатов поиска необходимо осуществить GET-запрос с параметрами, в качестве параметров указать извлеченные ключевые слова. При необходимости в параметрах можно указать дополнительные условия, например, огранить поиск определенными типами результатов (книги, ссылки на pdf файлы и т.д.) или выбрать параметр, по которому нужно сортировать результаты поиска. Далее необходимо осуществить синтаксический разбор полученной web-страницы и извлечь из нее ссылки на статьи.

Разрабатываемая система позволяет автоматизировать процесс поиска документов, релевантных содержанию заданного текста, что значительно снижает трудоемкость данного процесса и временные затраты.

Библиографический список

1. CleverStat.Sitecontentanalyzer [Электронныйресурс] URL: http://www.cleverstat.com/ru/sca-website-analysis-software-index.htm (дата обращения 13.03.2016).
2. Microsystems.Textanalyst[Электронныйресурс] URL: http://libots.sourceforge.net (дата обращения 05.03.2016).
3. Tesuçk.Веб-сервис Tesuçk[Электронныйресурс] URL: http://tesuck.eveel.ru (дата обращения 11.03.2016).
4. *EvansD.A., Lefferts R.G.*Clarit-trec experiments // Information processing &management. 1995. Vol. 31, no. 3. P. 385–395.
5. Google Scholar.Академия Google [Электронный ресурс] URL: https://scholar.google.ru (дата обращения 13.03.2016).

RELEVANT DOCUMENTS SEARCH SYSTEM

Polushin Gleb V.

National Research University Higher School of Economics,
st. Studencheskaya, 38, Perm, Russia, 614070, polushin.gleb@mail.ru

The article is dedicated to the relevant documents search process automation. Relevant literature search is a non-trivial task, as it requires time and imposes certain requirements to the reader. The paper proposes methods to minimize efforts required to find relevant literatureand to automatize the search process. The analysis of existing software solutions for extracting keywords from texts in Russian is given. Method of extracting key phrases fromRussian text using metrics TF-IDF is described. Method of automatic search of relevant documents with the extracted keywords using a free search engine for full texts of scientific publications of all formats and disciplines Google Scholar is also described. In practice, the system can be used by students, teachers and all who are engaged in scientific activities.

Key words: keywords extraction, automatic search, text processing.