УДК: 004.912

Разработка системы автоматического извлечения ключевых слов из текста

Р.Р. Гайнанов

Пермский национальный исследовательский политехнический университет

Работа посвящена такой области автоматической обработки текста как извлечение ключевых слов. Ключевое слово (КС, англ. keyword) рассматривается как слово или словосочетание из текста документа для передачи семантики документа в совокупности с другими ключевыми словами. В работе предлагается и поясняется подход с использованием словарей по предметным областям (domain dictionary). Проводятся эксперименты по извлечению ключевых слов из текстов по предметной области «Операционные системы» с помощью разработанной системы и сформированного словаря.

Ключевые слова: *NLP; Text Mining; Keywords Extraction; извлечение терминов; извлечение ключевых слов; обработка естественного языка; компьютерная лингвистика;*

Введение

В настоящее время объёмы и динамика информации, которая подлежит обработке в библиотечном деле, лексикографии и информационном поиске, делают особенно актуальной задачу автоматического извлечения ключевых слов и фраз, которые могут использоваться для создания и развития терминологических ресурсов, а также для эффективной обработки документов: индексирования, реферирования, кластеризация и классификации [1] [2].

Существует множество систем для автоматического извлечения ключевых слов, но лишь единицы из них имеет поддержку русского языка. Многие из их числа закрыты для общего использования и распространяются на коммерческой основе. Следовательно, существуют проблемы в выборе подходящего готового инструмента для извлечения ключевых слов из русскоязычных текстов.

В работе предлагается подход к извлечению ключевых слов из совокупности токенов, основанный на использовании словарей предметных областей (domain dictionary)[3], в отличие от широко распространенного статистического подхода [4]. В данном подходе ключевое внимание уделяется формированию словаря, в основу которого должны быть включены тщательно отобранные экспертом термины по предметной области. Использование словаря позволит исключить малозначащие слова среди набора ключевых слов.

Описание используемого подхода

Сам процесс извлечения ключевых слов из текста состоит из 4 этапов (рис. 1). После получения текста документа (*Текст*), он должен быть подвергнут процедуре токенизации, которая позволяет получить массив токенов – слов из реального контекста в том порядке в каком они идут в тексте (*Т-текст*). Для уменьшения избыточности получившегося массива производится исключение стоп-слов и знаков пунктуации (*Т'-текст*). Под стоп-словами (англ. stop-words) понимаются слова, встречающиеся практически во всех текстах и не несущие специальной смысловой нагрузки (предлоги, союзы, междометия и пр.).

Русский язык относится к группе флективных синтетических языков, то есть языков, в которых преобладает словообразование с использованием аффиксов, сочетающих сразу несколько грамматических значений, поэтому данный язык допускает использование алгоритмов стемминга. Стемминг – процесс нахождения основы слова для заданного исходного слова. В данной работе для этой процедуры используется распространенный открытый алгоритм Портера. Таким образом, анализируемый текст и словарь после разбиения на токены подвергается стеммингу. Полученный текст (*S-текст*) служит основой для извлечения ключевых слов. После происходит последовательный поиск словарных понятий в тексте и подсчет количества их повторений. И последним шагом производится отбор полученных понятий в список ключевых слов документа (*КС*).

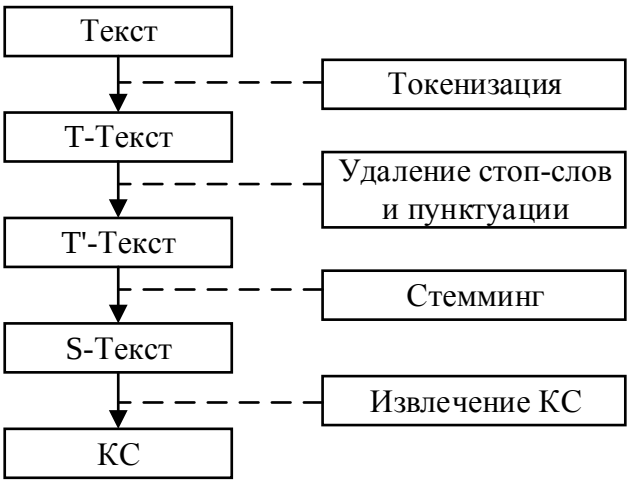

Рисунок 1 – Процесс извлечения ключевых слов

Первые четыре операции являются распространёнными в компьютерной обработке текста и существует множество инструментов по их реализации. Отдельно стоит остановиться на последней процедуре по извлечению ключевых слов и фраз на основе словаря из представления в виде S-текста.

На примере простого предложения: «*Процессы — это одна из самых старых и наиболее важных абстракций, присущих операционной системе*» после выполнения процедур токенизации и стемминга было получено следующее представление S-текста: «*процесс, одн, сам, стар, важн, абстракц, присущ, операцион, систем*». Именно оно показано на рис. 2, где поясняется рассматриваемый метод извлечения ключевых слов.

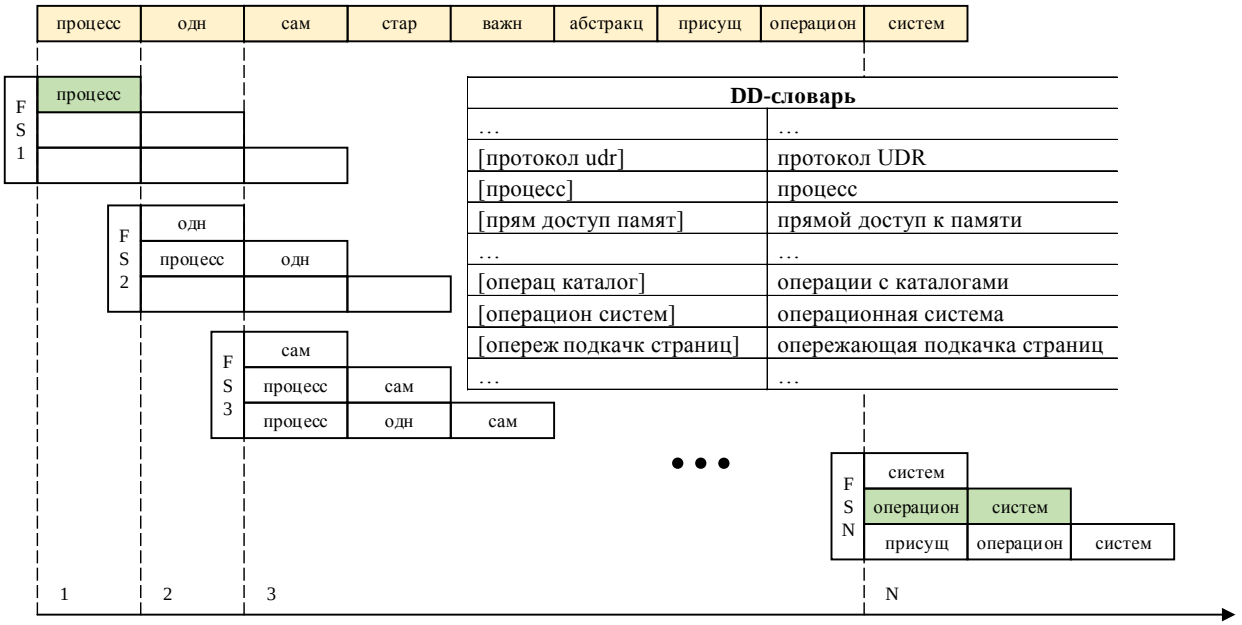

Рисунок 2 – Процедура извлечения ключевых слов из текста

Прежде чем работать с текстом, необходим составленный экспертом словарь по предметной области (*DD-словарь*). Он также подвергается процедурам токенизации и стемминга, в результате которых словарные понятия сохраняются в хэш-таблицу. В качестве ключа таблицы используется полученная основа слова, либо основы слов, разделенные пробелами, если понятие состоит из нескольких слов. Значение, на которое указывает ключ – это непосредственно само словарное понятие. Часть словаря также представлена на рис. 2.

После загрузки и обработки текста (получен массив *S-Текст*) происходит извлечение ключевых слов. Для этого производится последовательное чтение элементов массива *S-Текст* и заполнение структуры *FSx*, где *x* – номер прочитанного элемента. В основе структуры FSx лежит многомерный массив, где нулевой уровень (нулевая позиция) структуры представлен одномерным массивом с длиной в одно слово, и на каждом последующем уровне длина данного одномерного массива увеличивается на один. Тем самым достигается возможность извлечения словосочетаний из

текста. Каждый новый прочитанный элемент добавляется в нулевой уровень структуры и сдвигает ранее записанные элементы на следующие уровни. После каждого чтения происходит поиск текущих значений структуры в словаре. Если значение найдено производится запоминание его позиции в структуре и если данное запомненное значение на следующем шаге не увеличилось, то, следовательно, это понятие необходимо добавить в набор ключевых слов. Таким образом происходит извлечение самых длинных найденных фраз из текста, даже если в фразе присутствуют другие ключевые слова.

Сравнение результатов

За более чем полувековой период своего существования операционные системы прошли сложный путь развития, находясь под постоянным влиянием успехов в области вычислительной техники и информационных технологий. Эта популярная область, в которой постоянно публикуется большое количество научных статей и материалов и содержащая большое количество специфических терминов, была взята за основу при разработке системы и предметного словаря.

Для проведения экспериментов по извлечению ключевых слов был составлен набор русскоязычных текстов, близких к предметной области операционных систем. В основу данных текстов легли материалы книги Эндрю Таненбаума «Современные операционные системы» последней редакции на русском и английском языках [5] [6]. Именно эта книга является отличным представителем выбранной области, о чем можно судить, к примеру, по количеству цитирования в Google Scholar.

Для сравнения с разрабатываемой системой были отобраны следующие системы: OpenCalais, Extractor и Семантическое зеркало. Ключевыми факторами при отборе аналогов были доступность системы в виде web-сервиса и возможность проверить работу без покупки системы, количество исследовательских работ, посвящённых системе, а также популярность соответствующих систем в современном IT–сообществе.

Сравнительный анализ разработанной системы с аналогичными системами проводился на основе главы 2.1 «Процессы», стр. 111-123 (в англ. варианте – chapter 2.1 «Processes», pp. 85-97) упомянутого ранее источника. Каждый текст содержит около 5 тыс. слов. Результаты работы систем в табл. 1, 2.

Таблица 1 – Ключевые слова, полученные в сторонних системах

№	OpenCalais (англ. яз.)	Extractor (англ. яз.)	Семантическое зеркало (рус. яз.)	Семантическое зеркало (англ. яз.)
1.	Scheduling	CPU	программу	program
2.	Process state	operating system	процессора	CPU
3.	Computer multitasking	waiting	работе	systems
4.	Load	disk	память	new
5.	Task Manager	multiprogramming	систем	not
6.	Nice	UNIX	электронной почты	program counter
7.	Thread	program counter	центрального процессора	operating systems
8.	Parent process		операционных систем	new process
9.	CPU time		создание процесса	create processes
10.	Process		дочерний процесс	process run
11.	Concurrent computing		счетчика команд	processes running
12.	Process management		системного вызова	system call
13.			новые процессы	CPU utilization
14.			процесс управления	create new process
15.			входные данные	process execute

Таблица 2 – Ключевые слова, полученные в разрабатываемой системе

№	Ключевые слова (рус. яз.)	Вес	Ключевые слова (англ. яз.)	Вес
1.	процесс	6,03	Process	6,56
2.	центральный процессор	2,11	Operating system	1,34
3.	операционная система	1,21	Program counter	0,91
4.	процесс дочерний	0,78	Memory	0,79
5.	счетчик команд	0,72	Interrupt	0,67
6.	прерывание	0,60	System call	0,61
7.	системный вызов	0,48	UNIX	0,46
8.	создание процесса	0,48	Windows	0,39
9.	файл	0,42	Disk	0,39
10.	процесс родительский	0,42	Scheduler	0,36
11.	UNIX процессы	0,36	Multiprogramming	0,30
12.	адресное пространство	0,36	Address space	0,30
13.	последовательный процесс	0,30	Operating system design	0,27
14.	таблица процессов	0,30	Signal	0,24
15.	состояние процесса	0,30	Stack pointer	0,24
16.	иерархия процессов	0,30	Child process	0,24
17.	диск	0,27	Process table	0,24
18.	Windows	0,24	Process termination	0,24
19.	UNIX	0,24	System process	0,24
20.	событие	0,24	Processor	0,18

В результате в разрабатываемой системе для текстов на русском и английском языке был получен набор из 85 и 99 ключевых понятий соответственно, что составило около 3% от общего размера текста. В таблице 2 представлены первые 30 элементов данных наборов, ранжированных по весам. Для вычисления веса ключевого слова ($W(kw)$) использовалась формула (1).

$$W(kw) = \frac{Nw(kw) \cdot Freq(kw)}{N_t} \cdot 100\%, \quad (1)$$

где kw – ключевое слово;

$Nw(kw)$ – количество слов в kw;

$Freq(kw)$ – количество повторений kw в тексте;

N_t – размер S-текста.

Для оценки качества работы системы по извлечению ключевых слов применяются различные оценки, основанные на анализе результатов работы системы. При этом "идеальным" алгоритмом считается тот, для которого выводы, сделанные системой, согласуются с мнением оценивающих экспертов.

Таким образом, аналогичная цель по извлечению ключевых слов из текста на русском языке была поставлена для эксперта. Экспертом был составлен набор из 20 ключевых слов (табл. 3), принятый в дальнейшем за эталонный, чтобы произвести сравнение с наборами, полученными программно.

Таблица 3 – Набор ключевых слов, полученный экспертом

№	Ключевые слова	№	Ключевые слова
1.	процесс	11.	прерывание
2.	создание процесса	12.	вектор прерывания
3.	блокировка процесса	13.	операционная система
4.	завершение процесса	14.	многозадачность
5.	родительский процесс	15.	псевдопараллелизм
6.	дочерний процесс	16.	центральный процессор
7.	дескриптор процесса	17.	UNIX
8.	таблица процессов	18.	Windows
9.	адресное пространство	19.	системный вызов
10.	планирование	20.	демон

Большинство метрик, применяемых в современной оценке информационного поиска, могут быть применены и для оценки результатов систем извлечения ключевых слов. Данные метрики основываются на отношении релевантности. Обсуждение самого понятия релевантности выходит за рамки данной работы, и будет использоваться в сравнении с экспертным набором. Метрики для неупорядоченного набора ключевых слов основаны на бинарной классификации «релевантен/не релевантен» по отношению к выбранному тексту. Под выражением «ключевое слово релевантно» понимается, что данное слово имеется в наборе ключевых слов, полученных экспертом. Наиболее часто рассчитываемые метрики с точки зрения РОМИП – это полнота и точность [7]. Именно они будут использоваться в данной работе.

Полнота (recall) вычисляется как отношение найденных релевантных ключевых слов к общему количеству релевантных ключевых слов. Полнота характеризует способность системы извлекать релевантные ключевые слова, но не учитывает количество нерелевантных документов, выдаваемых пользователю. Например, если полнота равна 50%, то это значит, что половина релевантных ключевых слов системой не найдена.

Точность (precision) вычисляется как отношение найденных релевантных ключевых слов к общему количеству найденных ключевых слов. Точность характеризует способность системы выдавать в списке результатов только релевантные ключевые слова. Например, если точность равна 50%, то это значит, что среди найденных ключевых слов половина релевантных и половина – нерелевантных.

Таким образом, для полученных результатов разрабатываемой системы и сторонних систем были рассчитаны данные метрики и представлены в табл. 4.

Таблица 4 – Вычисление метрик качества наборов ключевых слов

Система	Язык	Количество КС	Полнота (R)	Точность (P)	F-мера (F)
Экспертные КС	Рус.	20	1	1	1
OpenCalais	Англ.	12	0,2 (4/20)	0,3 (4/12)	0,24
Extractor	Англ.	7	0,2 (4/20)	0,6 (4/7)	0,30
Семантическое зеркало	Рус.	15	0,25 (5/20)	0,3 (5/15)	0,27
КС с использованием словаря (топ-20)	Рус.	20	0,6 (12/20)	0,6 (12/20)	0,60
КС с использованием словаря	Рус.	85	0,9 (18/20)	0,2 (18/85)	0,33

Хорошей мерой для совместной оценки точности и полноты является сбалансированная F-мера, которая определяется следующим уравнением: $F = 2RP/(R + P)$. Максимальный показатель F достигается у набора из 20 ключевых слов полученных системой с использованием словаря. У другой системы, работающей с русским языком данный показатель сильно ниже.

Данные эксперименты и полученные результаты позволяют показать, что при использовании правильно составленного словаря по предметной области можно добиться увеличения показателей

полноты и точности, а, следовательно, и качества извлекаемых ключевых слов из русскоязычных документов по сравнению с действующими системами.

Заключение

Анализ состояния вопроса показал, что в данный момент на рынке присутствует дефицит в качественных системах для извлечения ключевых слов на русском языке. Поэтому особое внимание уделяется текстам на русском языке. Подход, основанный на специализированных словарях предметных областей, позволяет работать с текстами на разных языках. Также с помощью словарей становится возможным извлечение многословных понятий в той в форме, в которой они добавлены в словарь.

В результате данной работы был разработан прототип системы, производящей обработку загруженных словарей и извлечение ключевых слов из текстовых документов. На нем были получены экспериментальные результаты для текстов по предметной области «Операционные системы» и на основе экспертного набора им была дана оценка, которая позволила судить об увеличении качества извлекаемых ключевых слов по сравнению со сторонними аналогичными системами.

Литература

1. Bracewell, D. B., Ren F. Multilingual Single Document Keyword Extraction for Information Retrieval. Proceedings of NLP-KE, 2005, pp. 517-522.
2. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие. М.: МИЭМ, 2011. 272 с.
3. Dictionary Based Annotation at Scale with Spark, SolrTextTagger and OpenNLP [Электронный ресурс] / Sujit Pal // Spark Summit 2015. Europe. – URL: https://spark-summit.org/eu-2015/events/dictionary-based-annotation-at-scale-with-spark-solrtexttagger-and-opennlp.
4. Dostal M. Automatic Keyphrase Extraction Based on NLP and Statistical Methods. Proceedings of the Dateso 2011: Annual International Workshop on Databases, Texts, Specifications and Objects. Pisek, Czech Republic, 2011, pp. 140-145.
5. Таненбаум Э. С., Бос Х. Современные операционные системы. 4-е изд. – СПб.: Питер, 2015. – 1120 с.: ил.
6. Tanenbaum A. S., Bos H. Modern operating systems. – Prentice Hall Press, 2014.
7. Агеев, М. Приложение А. Официальные метрики РОМИП 2010 / М. Агеев, И. Кураленок, И. Некрестьянов // Труды РОМИП'2010. СПб.: Изд-во НУ ЦСИ. –2010. – с. 172-187.

Сведения об авторе:

1. ***Гайнанов Руслан Рамилевич***, магистрант кафедры Информационных технологий и автоматизированных систем Пермского национального исследовательского политехнического университета.

Область научных интересов: системы автоматической обработки текста, распределенные и облачные вычисления, базы данных и распределенные системы хранения.

Контактный e-mail: ruslan.r.gainanov@gmail.com