Разработка, оценка и использование алгоритма сегментации слов для систем мониторинга национальных интернет-ресурсов

© Березин Б.А. © Ландэ Д.В.
Институт проблем регистрации информации Национальной академии наук Украины, Киев, Украина
bberua@ukr.net dwlande@gmail.com
© Павленко О.Ю.
Открытый международный университет развития человека "Украина", Киев, Украина
asd97456@gmail.com

Аннотация

Растущее количество информационных ресурсов, представленных в Интернете, ведет к необходимости развития поисковых систем для доступа к ним. В то же время, растет доля и важность мировых веб-ресурсов, представленных на языках, поиск в которых требует определения границ слов (китайский, японский, тайский и др.), в текстах которых не принято разделение слов. В данной статье рассмотрены особенности алгоритмов сегментации слов из таких текстов. Сегментация необходима для использования традиционных поисковых систем с индексированными словами (типа Google, Bing) при доступе к китайским, японским, тайским и подобным веб-ресурсам. Отмечаются две основные модели — статистические и использующие словарь. Для моделей со словарем отмечается вариант алгоритма с максимальным совпадением (maximal matching), для которого есть модификации – Forward Maximal Matching (FMM) и Backward Maximal Matching (BMM), в зависимости от направления обработки текста. Второй вариант для моделей со словарем - это алгоритм, который находит сегментацию с минимальным количеством слов.

В работе представлен новый алгоритм сегментации слов на основе модифицированного волнового алгоритма. Алгоритм учитывает особенности входных данных и построен так, что необходимые вычисления выполняются за один проход. Это снижает его вычислительную сложность. Приведено описание и псевдокод алгоритма сегментации слов. Показан пример разбиения входной строки на английском языке на слова, представления ее в виде графа и нахождения кратчайшего пути.

Для оценки качества сегментации приведен метод EDWS (Edit Distance of the Word Separator). При этом использовался специальный инструментарий для оценки сегментации китайских слов с тестовым корпусом на основе новостных текстов. Получены оценки качества сегментации слов для предложенного алгоритма (на основе поиска кратчайшего пути) и ряда других известных сегментаторов. Приведен пример сегментации новостного текста на русском языке. Показаны возможности использования разработанного алгоритма в задачах поиска информации в национальных сементах сети Интернет. Реализация алгоритма сегментации слов используется при создании обобщенной модели предметной области на базе мониторинга ресурсов китайского сегмента Интернет.

Рост количества информационных ресурсов китайского сегмента сети Интернет обуславливает необходимость создания глобальных информационно-поисковых систем. Для реализации поисковых индексов таких систем необходима быстрая, точная и полная сегментация слов из текстов. Полученные оценки качества сегментации с помощью предложенного алгоритма при формировании индекса поисковой системы свидетельствуют о возможности его использования для информационных ресурсов китайского сегмента сети Интернет.

Ключевые слова: сегментация слов, качество сегментации слов, мониторинг веб-ресурсов, поиск кратчайшего пути, волновой алгоритм.

1 Актуальность и анализ публикаций

В работах [1, 2] отмечаются такие особенности китайского Интернет-пространства как высокие темпы роста веб-ресурсов и числа пользователей; наличие собственных социальных сетей и поисковой системы Baidu, ориентированной на китайский язык (существенной проблемой является применение латиницы и кириллических кодов) и покрывающей значительную часть веб-ресурсов китайского сегмента Интернет. В этих работах также показаны подходы к построению систем мониторинга национальных Интернет-ресурсов, где актуальной является сегментация слов при формирования индекса поисковой системы.

В ранней работе [3], рассматривается портирование систем поиска и извлечения информации в сегмент ресурсов, представленных на китайском и японском языках. Отмечается важность индексирования китайских символов и проблемы автоматической сегментации. В [4] отмечается, что для решения проблемы сегментации китайского текста используются три основных способа: словарный (обычно с применением алгоритма максимального соответствия), статистический и комбинированный, сочетающий в себе оба предыдущих. При

этом, удаётся правильно сегментировать текст более чем на 90%. Работа [5] рассматривает сегментацию слов языка урду (который также не использует разделителей между словами) как важную проблему приложений обработки естественных языков (NLP). Рассматриваются технология совпадения с наиболее длинными словами из словаря (longest matching); технология максимального совпадения (maximum matching) и методы статистической сегментации.

Данная работа посвящена реализации модели алгоритма сегментации слов (АСС) для формирования индекса поисковой системы. Показаны варианты АСС, предложена реализация моделей алгоритма сегментации, получена оценка качества сегментации. Реализованные АСС используются авторами для создания обобщенной модели предметной области на базе мониторинга ресурсов китайского сегмента Интернет.

2 Особенности моделей алгоритма сегментации слов

В работах, посвященных сегментации слов, выделяются две основные модели — статистические и использующие словарь (правила, списки слов). Оценка этих двух подходов к сегментации выполнена в работе [6]. Для моделей на основе словаря должен быть доступен предопределенный словарь. При этом отмечается вариант алгоритма с максимальным совпадением (maximal matching), для которого есть модификации - forward maximal matching (FMM) and backward maximal matching (BMM), в зависимости от направления обработки текста. Второй вариант для алгоритма со словарем - это алгоритм, который находит сегментацию с минимальным количеством слов shortest path (SP).

Для моделей на основе словаря предполагается наличие списка слов, каждое из которых связано с оценкой вероятности того, что оно является истинным словом. Пусть

$$W = \{\{w_i, g(w_i)\}_{i=1,\ldots,n}\}$$

будет таким списком, где w_i является кандидатом на слово, а также $g(w_i)$ его функция качества. Алгоритм прямого максимального соответствия FMM обрабатывает текст T для вывода лучшего текущего слова w^*многократно с T=t* для каждого цикла, таким образом

$$\{w^*, t^*\} = \underset{wt=T}{\operatorname{argmax}}\, g(w)$$

с каждым $\{w, g(w)\} \in W$.

Алгоритм сегментации на основе кратчайшего пути [6,7] использует предположение о том, что правильная сегментация должна максимизировать длины всех слов или минимизировать общее количество слов. Для предложения *S* из *m* символов $\{c_1, c_2, \ldots, c_m\}$ лучшее сегментированное предложение S^* из n^* слов $\{w_1^*, w_2^*, \ldots, w_{n^*}^*\}$

$$S^* = \underset{w_1 \cdots w_i \cdots w_n = T}{\operatorname{argmin}} n.$$

Эта задача оптимизации преобразуется в задачу нахождения кратчайшего пути для направленного нециклического графа.

3 Разработка алгоритма сегментации

Для анализа АСС были рассмотрены различные реализации моделей сегментатора. С помощью языка Perl на основе алгоритма максимального соответствия FMM было разработано соответствующее программное обеспечение. Также рассматривалась предложенная в [8] реализация сегментатора на основе алгоритма максимального соответствия. При поиске слов он пытается использовать самое длинное возможне слово. Этот простой алгоритм достаточно эффективен при большом словаре. Особенностью алгоритма являются элементы средств идентификации и извлечения китайских именованных объектов. В реализации алгоритма на Perl описаны такие объекты, как числа, ASCII коды, китайские фамилии и т. п. и предусмотрены процедуры для их извлечения. Более подробно система извлечения китайских именованных объектов рассмотрена в [9], где приводятся несколько видов правил для отнесения слов к именованным объектам. Реализация сегментатора [8] была адаптирована и использована для формирования индекса поисковой системы при работе с новостными, научно-техническими и др. веб-ресурсами китайского сегмента Интернет.

В данной работе, для анализа и усовершенствования модели АСС предложен и разработан алгоритм с поиском кратчайшего пути в графе [10]. Его реализация выполнена на языке Perl. Предложенный алгоритм сегментации состоит из трех частей: Формирование таблицы Слов, Формирование таблицы Шаг-Позиция, Формирование массива сегментированных слов и вывод в файл. В первой части программы создается массив, каждая строка которого соответствует символу входной строки. При нахождении множества слов, на которые может быть разбита входная строка, для каждой входной буквы анализируются возможные подстроки,

начинающиеся с данной буквы, длиной от 1 до n. (n – максимально возможная длина слова, зависит от языка. Для китайского, в словаре можно найти слова до 5-6 иероглифов, для русского до 18-20 букв и т.д.). Если для анализируемой подстроки находится соответствующее слово в словаре, то такое слово используется в разбиении. Построенная таким образом таблица содержит множество слов, на которые может быть разбита входная строка при данном объеме словаря. В таблице 1 показано разбиение, полученное для входной строки IWORKATTHERESEARCHINSTITUTE (в словаре найдены слова и часто используемые сокращения).

Множество слов, на которые разбивается входная строка, может быть представлено направленным графом. Буквы входной строки являются вершинами такого графа, а слова разбиения, получаемые для каждой буквы, представляют ребра, исходящие из соответствующих вершин. На рис. 1 приведено представление разбиения входной строки IWORKATTHERESEARCHINSTITUTE на слова в виде графа. При таком представлении, выбор слов правильного разбиения входной строки может рассматриваться как поиск минимального пути в графе. Различные варианты волнового алгоритма, наряду с другими методами нахождения минимального пути в графе, могут использоваться для решения этой задачи.

Таблица 1. Разбиение, полученное для входной строки IWORKATTHERESEARCHINSTITUTE

Входная строка	**Найденные в словаре слова разбиения**				
i	**i**				
w	**work**				
o	**or**				
r					
k					
a	**a**	**at**			
t					
t	**th**	**the**	**there**	**theres**	
h	**he**	**her**	**here**		
e	**ere**				
r	**re**	**res**	**research**		
e	**es**				
s	**se**	**sea**	**sear**	**search**	
e	**ear**				
a	**a**	**ar**	**arc**	**arch**	
r					
c	**ch**	**chi**	**chin**	**chins**	
h	**hi**				
i	**i**	**in**	**ins**	**inst**	**institute**
n	**ns**				
s	**st**				
t	**ti**	**tit**			
i	**i**	**it**			
t	**tu**	**tut**			
u	**ut**				
t					
e	**e**				

Во второй части программы, в соответствии с волновым алгоритмом, с помощью таблицы Слов формируется таблица ШагПозиция, с помощью которой выполняется разметка вершин графа, определяется их расстояние (соответствующие количеству ребер в случае невзвешенного графа) от начальной вершины (первой буквы входной строки). Для этого, на каждом шаге (волне) алгоритма определяется множество соседних, достижимых в данный момент вершин. Т.е., таблица ШагПозиция содержит множество вершин графа, которые достижимы при анализе текущего входного символа (текущей строки таблицы Слов). А также содержит позиции начала слов, на которые разбивается входная строка при сегментации. В целом, программа может рассматриваться как модифицированный волновой алгоритм нахождения минимального пути в графе.

На третьем этапе алгоритма, начиная с последней вершины, на основе расстояний до предшествующих вершин, производится выбор ребер в графе, составляющих минимальный путь, т.е. выбор минимального количества слов, на которые может быть разбита, сегментирована входная строка. На этом этапе, при выборе слов, для улучшения качества сегментации, кроме расстояний могут также учитываться другие особенности языков, например частота использования слов в текстах на разных языках и т.п. Т.е., в этой части программы, на основе таблицы ШагПозиция и входной строки символов заполняется массив Сегментированных Слов. Затем слова из сформированного массива Сементированных Слов выводятся в выходной файл программы.

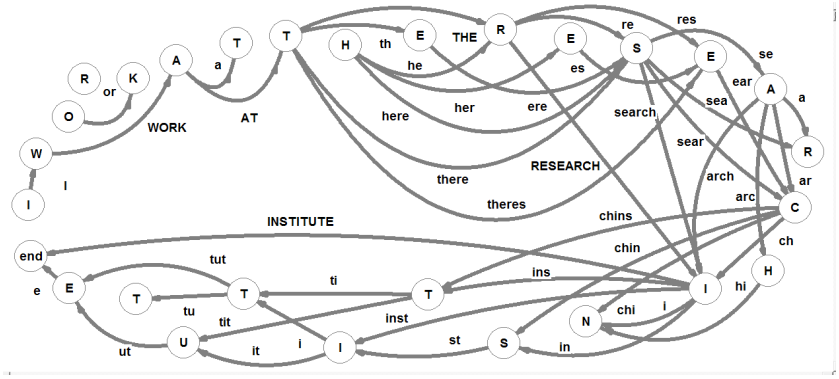

Рис. 1. Представление разбиения входной строки IWORKATTHERESEARCHINSTITUTE на слова в виде графа.

4 Оценка и использование моделей сегментации

Для оценки качества сегментации и возможности использования реализованных вариантов ACC проводилось тестирование алгоритма максимального соответствия FMM и алгоритма с поиском кратчайшего пути в графе, а также других, эксплуатируемых в сети Интернет сегментаторов. Аналогично работам [11, 12], для оценки сегментаторов использовался инструментальный набор для оценки сегментации китайских слов https://web.archive.org/web/20100702191858fw_/
http://projectile.sv.cmu.edu/research/public/tools/segmentation/eval/index.htm. В состав набора входят программные средства и тестовый корпус. При оценке сегментаторов использовался тестовый корпус из инструментального набора на основе китайских новостных текстов (данные проекта the Chinese Treebank, с китайскими текстами из Xinhua news, Sinorama news magazine, and Hong Kong News [11]). Кроме того, в данной работе был подготовлен тестовый корпус объемом 10 тысяч слов на основе частотного китайского словаря.

Метод оценки, который использовался для тестирования сегментаторов, называется Edit Distance of the Word Separator (расстояние при редактировании разделения слов, EDWS).

Пусть $C_1C_2....C_iC_{i+1}...C_n$. представляет предложение без сегментации. Сегментатор разделяет это предложение на последовательность слов путем вставки разделителя ***S*** между символами, в результате предложение может иметь, например, такой вид: $C_1C_2\boldsymbol{S}C_3C_4C_5\boldsymbol{S}C_6C_7C_8C_9\boldsymbol{S}C_{10}\boldsymbol{S}....C_iC_{i+1}...\boldsymbol{S}C_n$. Разные программы сегментации могут размещать разделители в разных позициях одного и того же предложения. Расстояние при редактировании разделения слов (EDWS) определяет, сколько операций редактирования (вставки, удаления и совпадения) необходимы для модификации автоматически полученной сегментации текста до стандартной сегментации того же текста, полученной вручную, экспертами. Точность (precision) и полнота (recall) являются показателями, которые используются при оценке большей части алгоритмов извлечения информации. Для оценки сегментации эти показатели могут быть определены так:

Precision= (число совпадений) / (число разделителей в полученном тексте) *(1)*

Recall=(число совпадений) / (число разделителей в стандартном тексте) *(2)*

F=2*Precision*Recall / (Precision+Recall) *(3)*

В таблице 2 в столбце 1 показаны виды сегментаторов, для которых проводилось тестирование. В столбце 2 — тип корпуса (новостной, из инструментального набора, или словарный, отобранный из словаря), использованный при тестировании. В 3-5 столбцах таблицы приведены количества операций редактирования (совпадения, вставки, удаления), которые необходимо было выполнить для модификации текста после сегментатора в стандартный текст. В последних, 6-8 столбцах представлены значения показателей, рассчитанные по формулам (1) - (3) при проведении тестирования. Для алгоритма Е. Петерсона тестирование было проведено с использованием словарей двух размеров - 348982 слов (один из словарей сегментатора Jieba) и 119804 слов (словарь WordList).

Кроме тестирования алгоритмов сегментации на основе китайских текстов, для алгоритма максимального соответствия FMM и алгоритма с поиском кратчайшего пути в графе для оценки разбиения проводилось также тестирование с использованием русских и английских текстов и словарей. Оценивался процент слов, сегментированных с ошибками для массивов данных из различных видов информационных ресурсов - новостных сообщений, технических и художественных текстов. При этом использовались два вида словарей - словарь русского языка и словарь, сформированный путем накопления новостных сообщений.

Реализованные и адаптированные ACC использовались для мониторинга китайских Интернет-ресурсов. Интерфейс и результаты отработки запроса на мониторинг при использовании сегментатора на основе адаптированного алгоритма Е. Петерсона [8] в китайской социальной сети Weibo показаны на рис. 2. Кроме того, в рассматриваемой модели реализованы средства поиска документов в получаемых результатах мониторинга.

Таблица 2 – Результаты тестирования

Вид сегмента-тора	Тестовый корпус	Сов.	Вст.	Удл.	Precision	Recall	F-1
1	2	3	4	5	6	7	8
Jieba	новостной	19581	2567	1233	94.08%	88.41%	0.91
	словарный	9966	33	365	96.47%	99.67%	0.98
Online	новостной	21334	814	1911	91.78%	96.32%	0.94
	словарный	9686	313	3508	73.41%	96.87%	0.84
Алгоритм с поиском кратчай-шего пути в графе (словарь 348982 слов)	новостной	19313	2835	1699	91.91%	87.20%	0.89
	словарный	9988	11	464	95.56%	99.89%	0.98
Алгоритм Петерсона (словарь 348982 слов)	новостной	20673	1475	1073	95.07%	93.34%	0.94
	словарный	9882	117	2940	77.07%	98.83%	0.87
Алгоритм Петерсона (словарь 119804 слов)	новостной	20524	1624	1452	93.39%	92.67%	0.93
	словарный	9696	303	4850	66.66%	96.97%	0.79
Ir-Segmenter	новостной	21345	803	3145	87.16%	96.37%	0.92

Рис. 2. Отработка запроса на мониторинг в китайской социальной сети Weibo

5 Выводы

Показана актуальность задачи сегментации слов при формировании индекса поисковых систем в связи с ростом ресурсов китайского и др. сегментов Интернет. Приведены варианты ACC, которые могут быть использованы для формирования индекса поисковой системы, показана применимость моделей на основе словаря.

Рассмотрены модели реализации FMM ACC на основе словаря. Предложен алгоритм сегментации с поиском кратчайшего пути в графе и разработано программное обеспечение.

Получены оценки качества сегментации и результаты использования модели ACC при формировании индекса поисковой системы для мониторинга веб-ресурсов китайского сегмента Интернет, которые показывают возможность использования алгоритма при достаточном объеме словаря.

Литература

1. Ландэ Д.В. Обзор особенностей и возможности контент-мониторинга национального сегмента сети Интернет / Д.В. Ландэ, Б.А. Березин, В.А. Додонов // Реєстрація, зберігання і обробка даних, - 2016. - Т. 18, - N 3. - С. 20-38.
2. 2.Ландэ Д., Березин Б., Павленко О. Построение модели информационного сервиса на базе национального сегмента Интернет // Информационные технологии и безопасность. Материалы XVI Международной научно-практической конференции ИТБ-2016. - К.: ИПРИ НАН Украины, 2017. - С. 48-57.
3. 3.Boisen, S. Chinese information extraction and retrieval / S. Boisen, M. Crystal, E. Peterson, R. Weischedel, J. Broglio, J. Callan, M. E. Okurowski // Proceedings of a workshop on held at Vienna, Virginia. Association for Computational Linguistics, - 1996. - P. 109-119.
4. 4.Загибалов Т.Е. Автоматический анализ текстов на китайском языке. Проблема выбора базовой единицы // Труды международной конференции " Диалог", - 2005. – С. 31-37.

5. 5.Durrani, N., Hussain, S. Urdu word segmentation // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010. –P. 528-536.
6. Zhao H., Utiyama M., Sumita E., Lu B. L. An empirical study on word segmentation for chinese machine translation // International Conference on Intelligent Text Processing and Computational Linguistics. Springer Berlin Heidelberg, 2013. - P. 248-263.
7. Jia Z., Wang P., Zhao H. Graph model for Chinese spell checking // Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing (SIGHAN'13), 2013. - P. 88-92.
8. [Электронный ресурс]. — Режим доступа: http://www.mandarintools.com/segmenter.html. Peterson Erik. — Название с экрана.
9. Peterson Erik. A Chinese named entity extraction system // Proceedings of the 8th Annual Conference of the International Association of Chinese Linguistics, Melbourne, Australia. 1999. – P. 47-58.
10. Ландэ Д.В., Березин Б.А., Павленко О.Ю. Разработка алгоритма сегментации слов для систем мониторинга национальных интернет-ресурсов //Міжнародна науково-практична конференція "Інтелектуальні технології лінгвістичного аналізу": Тези доповідей.- Київ: НАУ, 2017. - С. 11.
11. Fung R., Bigi B. (2015, October). Automatic word segmentation for spoken Cantonese // Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2015 International Conference IEEE. -P. 196-201.
12. Chea V., Thu Y.K., Ding C., Utiyama M. Khmer word segmentation using conditional random fields // Khmer Natural Language Processing, 2015. – P. 62-69.

Development, Evaluation and Usage of Word Segmentation Algorithm for National Internet Resources Monitoring Systems

© Boris A. Berezin © Dmitry V. Lande
Institute for information recording of National academy of sciences of Ukraine,
Kyiv, Ukraine
bberua@ukr.net dwlande@gmail.com
© Oleh Y. Pavlenko
Open International University of Human Development "Ukraine",
Kyiv, Ukraine
asd97456@gmail.com

Abstract

The growing amount of information resources in the Internet leads to a need for search engines development. At the same time, increasing the part and importance of the world's web-resources, represented in Chinese, Japanese, Thai, etc. Search in that web-resources requires to define boundaries of words due to a lack of separation symbols in their texts. In this article, the features of word segmentation algorithms from such texts are considered. Segmentation is essential for using the traditional search engines with indexed words (such as Google, Bing) for access to the Chinese, Japanese, Thai and similar web-resources. There are two main models – statistical and using a dictionary. Also there is a variant of the algorithm with a maximal matching that have modifications – Forward Maximal Matching (FMM) and Backward Maximal Matching (BMM), depending on the text processing direction, for models that using a dictionary. The second variant is for this models is an algorithm that finds a segmentation with a minimum amount of words.

The paper presents a new algorithm for words segmenting based on a modified wave algorithm. The algorithm takes into account the features of input data and was developed for performing essential calculations in a single pass. It reduces the computational complexity. It is given a description and a pseudocode of the word segmentation algorithm. An example of a splitting of an English input string into words, a representing the string in a graph form and a finding the shortest path was introduced.

For assessment the quality of segmentation the EDWS (Edit Distance of the Word Separator) method was introduced. In this case, a special tool was used for assessment the quality of Chinese words segmentation with a test corpora based on the texts of news. Assessments the quality of words segmentation for the proposed algorithm (based on the shortest path search) and some other known segmentators are obtained. An example of segmentation of a news text in Russian is given. Opportunities of the developed algorithm for information search in national segments of the Internet are shown. The implementation of the word segmentation algorithm is using for creating a generalized domain model based on monitoring of the Chinese Internet segment resources.

An increasing of quantity of information resources of the Chinese Internet segment causes a necessity of global information retrieval systems creation. For implementation of search indexes for such systems, fast, accurate and complete segmentation of words from texts is required. Obtained with the proposed algorithm segmentation quality assessments indicate opportunity of using it for the Chinese Internet-segment resources.

Keywords: words segmentation, word segmentation quality, web resources monitoring, shortest path search, wave algorithm.