Автоматическое составление обзорных рефератов новостных сюжетов♣

© С.Д. Тарасов

Балтийский Государственный Технический Университет им. Д.Ф.Устинова «ВОЕНМЕХ»
tarasov_sd@mail.ru

Аннотация

Работа посвящена одной из актуальных проблем автоматического реферирования – составлению обзорных рефератов по набору документов. Рассмотрен новый на сегодняшний день алгоритм ранжирования связных структур (Manifold Ranking Algorithm) применительно к автоматическому реферированию новостных сюжетов. Алгоритм позволяет учитывать как зависимости между предложениями внутри одного документа, так и зависимости между всеми предложениями коллекции. Проведен анализ возможности использования алгоритма для русского языка. Построена пробная система автоматического реферирования. Приведены результаты работы системы. Сформулированы основные проблемы реализации системы и возможные методы их решения. Оценка качества работы системы произведена при помощи критерия ROUGE. Произведено сравнение результатов работы построенной системы с результатами в DUC 2003, DUC 2005.

1 Введение

Задача автоматизации реферирования текстовой информации на сегодняшний день остается очень актуальной, несмотря на огромное количество появившихся в последние годы публикаций. Это вызвано, в первую очередь, необходимостью в условиях постоянного роста информации знакомить специалистов и других заинтересованных людей с необходимыми им документами, представленными в сжатом виде, но с сохранением их смысла. В обзорной статье [1] описывается современное состояние в области автоматического реферирования, а также основные направления и пути развития. В традиционных методах реферирования чаще всего используются различные модификации подхода Г. Луна [9], известного с конца 50-х годов XX века, который заключается в отборе предложений с наибольшим весом для включения их в реферат, а также подходы, сочетающие традиционный подход с некоторыми новыми элементами. Вес предложения определяется как сумма частот, входящих в него значимых слов. В работе [2] описан метод, в котором в качестве значимых элементов выбираются не слова, а словосочетания. Развитие этого подхода есть в работе [3].

При формировании и показе сообщений новостных сюжетов приобретает актуальность задача составления обзорных рефератов (по некоторому набору документов), в которых были бы представлены все основные вопросы, затрагиваемые в каждом документе, но в обобщенном виде без повторений информации.

Составление обзорных рефератов относится к новым сферам применения автоматического реферирования, также как и получение одноязычных рефератов, охватывающих источники на разных языках, использование гибридных источников (например, статистической информации и сведений из баз данных), составление мультимедийных рефератов.

За рубежом в рамках конференций по проблемам автоматического аннотирования DUC (Document Understanding Conference) и текстового реферирования TSC (Text Summarization Challenge) данному направлению исследований придается очень большое значение. Автоматическое реферирование набора новостных сюжетов реализовано в таких крупных новостных ресурсах, как Google News (http://news.google.com/), NewsBlaster (http://www.newsblaster.com/), Yandex News (http://news.yandex.ru/).

Недавно разработанный алгоритм ранжирования связных структур (Manifold Ranking Algorithm) [4] может быть использован для ранжирования любых информационных примитивов: текстов, предложений, изображений, звуков. В этом случае любой вид информации должен быть представлен в векторном пространстве. В задачах автоматического аннотирования данный алгоритм может быть применен для ранжирования предложений набора документов и отбора наиболее значимых из них для включения в обзорный реферат. В задачах ранжирования результатов информационного поиска «отправной точкой»

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2008, Дубна, Россия, 2008.

алгоритма является запрос, и ранг предложений определяется как мера их «информационной близости» запросу. В задаче автоматического реферирования «отправной точкой» алгоритма можно считать «тему» кластера. В этом случае, правда, тема должна быть сформулирована максимально четко и подробно, т.к. алгоритм основан на анализе лексики и не может учитывать возможную семантическую близость лексически «удаленных» предложений. Принцип автоматического реферирования набора документов на основе алгоритма пространственного ранжирования подробно описан в [5]. Там же приведена довольно подробная оценка работы метода. Мне представилось интересным опробовать данный алгоритм для автоматического реферирования новостных сюжетов на русском языке, а также провести свои оценки качества реферирования. В процессе работы были выявлены некоторые недостатки метода, а также намечены основные возможные пути улучшения качества реферирования.

2 Алгоритм ранжирования связных структур

2.1 Обзор

Алгоритм Manifold Ranking позволяет описать связную структуру текста при помощи матриц. Изначально алгоритм предполагает выделение элементов (предложений) наиболее близких заданному (теме). Такая интерпретация характерна задаче информационного поиска. Для автоматического реферирования также выделяется набор предложений, наиболее близких заданной теме кластера, однако обязательным является применение алгоритма отсечения «похожих» предложений, что особенно актуально для многодокументного аннотирования.

Автоматическое реферирование набора документов с использованием алгоритма ранжирования связных структур состоит из двух этапов:

Вычисление ранга каждого предложения. Этим решается задача ранжирования всех предложений в соответствии с их «близостью» заданной теме кластера.

Применение алгоритма отсечения предложений, наиболее похожих на те, что уже попали в обзорный реферат. Этим решается задача исключения из обзорного реферата одинаковых или близких предложений.

В результате некоторое количество предложений с наибольшим рангом выбирается для результирующего реферата. Порядок следования предложений в общем случае никак не специфицируется подходом. Мной был реализован самый простейший алгоритм выборки предложений в порядке их относительного следования с приоритетом для более коротких предложений, что является естественным для русского языка. Строго говоря, вопрос связности полученного реферата является отдельной темой исследования. Некоторые методы решения представлены в [6]. Исходя из спецификации, алгоритм ранжирования связных структур оперирует двумя понятиями:

Информационная значимость: По заданному набору предложений $\overline{\mathrm{X}} = \{x_i \mid 1 \le i \le n\}$ и заданной теме T вычисляется вектор $\overline{f} = \{f_0 \cdots f_n\}$ информационной значимости каждого предложения $\overline{x_i}$. Информационная значимость предложения определяется как степень близости к заданной теме T. Предполагается, что тема кластера T наиболее полно отражает содержание набора документов и содержит наиболее полный набор лексики.

Информационная новизна: Для каждого предложения определяется его близость с другими предложениями набора. В итоге суммарный рейтинг, который определяет попадание предложения в обзорный реферат, рассчитывается с учетом, как информационной значимости предложения, так и его «информационной новизны».

2.2 Алгоритм

Алгоритм ранжирования связных структур является универсальным алгоритмом ранжирования объектов с учетом их внутренней связной структуры. Объекты должны быть представлены векторами в Евклидовом пространстве. В этом случае полагается, что «близость» двух объектов представленных векторами может быть вычислена, как Евклидова мера или скалярное произведение векторов. Целью алгоритма является упорядочить объекты, с учетом внутренних связей объектов между собой. Формально, связная структура объектов представляется как некий взвешенный граф, вершинами которого являются сами объекты, а в качестве весов дуг задаются евклидовы расстояния между объектами. В случае ранжирования предложений с целью отбора наиболее значимых из них для построения обзорного реферата алгоритм можно формализовать следующим образом:

[1] Задается набор структур (предложений) $\overline{\mathrm{X}} = \{\overline{x_i} \mid 1 \le i \le n\} \subset R^m$, где $\overline{x_0}$ – описание темы кластера T. Мы полагаем, что тема формулируется одним предложением.

[2] Вводится $f : \overline{X} \to R$ – отображение, которое ставит в соответствие каждой точке $\overline{x_i}(0 \le i \le n)$ значение ранга f_i. Мы можем рассматривать f как вектор $f = [f_0, f_1, .. f_n]^T$

[3] Задается вектор $y = [y_0, y_1, .. y_n]^T$. Согласно алгоритму $y_0 = 1$, т.к. $\overline{x_0}$ – тема кластера (в задачах информационного поиска $\overline{x_0}$ соответствует фразе поискового запроса), и $y_i = 1, i \in (1, n)$ для всех остальных предложений.

[4] Каждое предложение (объект) представляется в векторном пространстве следующим образом: $x_i = [tf_0, tf_1, ..tf_n]^T$, где tf_k – стандартная TF_ISF мера относительной важности терма t_k.

[5] Набор предложений представляет собой взвешенный граф с матрицей весов *W*. Для каждой пары $\overline{x_i}$ и $\overline{x_j}$ предложений вычисляется вес их «лексической близости» при помощи стандартной евклидовы меры. Таким образом:

$$W_{i,j} = Sim(\overline{x_i}, \overline{x_j}) \quad (1)$$

где

$$Sim(\overline{x_i}, \overline{x_j}) = \frac{\overline{x_i} \cdot \overline{x_j}}{\|x_i\| \cdot \|x_j\|} \quad (2)$$

Причем $W_{i,i} = 0$ для того, чтобы полученный граф не содержал циклов. Следует отметить, что полученная матрица весов является симметричной относительно своей главной диагонали.

Рассмотрим пример для случая с тремя предложениями-кандидатами, которые необходимо ранжировать.

Тема кластера: «Мама мыла раму»:

Таблица 1. Предложения кластера

№	Обозн.	Текст
1	X0	Первого октября мама мыла раму
2	X1	Мама мыла раму
3	X2	Мама мыла раму тряпкой

В результате морфологического разбора могут быть вычислены TF_ISF метрики для всех термов предложений (табл. 2). Предложение с номером 0 представляет собой тему кластера (x_0).

Таким образом, исходные данные представления предложений в евклидовом пространстве будут следующие:

$$x_0 = (0.33;0.33;0.00;0.00;0.33;0.00)$$
$$x_1 = (0.20;0.20;0.48;0.48;0.20;0.00)$$
$$x_0 = (0.33;0.33;0.00;0.00;0.33;0.00)$$
$$x_0 = (0.25;0.25;0.00;0.00;0.25;0.60)$$

Матрица весов в этом случае будет выглядеть следующим образом:

$$W = \begin{matrix} 0{,}000 & 0.457 & 1.000 & 0.587 \\ 0{,}457 & 0{,}000 & 0{,}457 & 0{,}268 \\ 1{,}000 & 0.457 & 0{,}000 & 0{,}587 \\ 0{,}587 & 0{,}268 & 0{,}587 & 0{,}000 \end{matrix}$$

Построенный граф может быть представлен как:

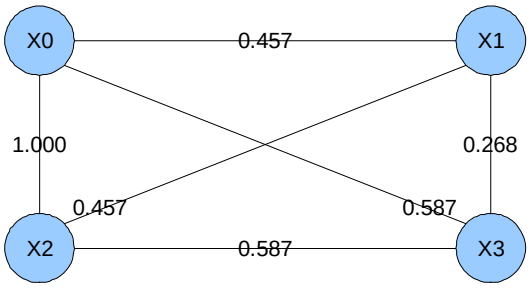

Рис. 1. Граф связности текста

Матрица весов подвергается симметричной нормализации

$$S = D^{-1/2} \cdot W \cdot D^{-1/2}, \quad (3)$$

где *D* – диагональная матрица, где $D_{i,i}$ равен сумме элементов i-ой строки матрицы *W*. Нормализация матрицы необходима, для того, чтобы итеративный алгоритм сходился.

$\overline{f}$ вычисляется как результат итеративного процесса:

$$\overline{f}(t+1) = \alpha \cdot S \cdot \overline{f}(t) + (1-\alpha) \cdot \overline{y} \quad (4)$$

Согласно теореме в [4] итеративный процесс сходится к f^*. Далее полагается, что f_i^* – полученный ранг предложения с номером i. Интуитивно, алгоритм заключается в постепенном распространении объектами своего ранга на смежные объекты-вершины. Таким образом, ранг f_i каждого предложения $\overline{x_i}$ вычисляется не только с учетом «близости» его к эталонному объекту (теме кластера T), но и с учетом связной структуры текста, т.е. ранг «распространяется» по графу с учетом весов связей структур.

Для приведенного выше примера итеративный процесс для вычисления ранга выглядит следующим образом:

$$f^0 = (0.00;0.00;0.00;0.00)$$
$$f^1 = (0.40;0.00;0.00;0.00)$$
$$f^2 = (0.40;0.09;0.12;0.09)$$
$$f^3 = (0.46;0.13;0.14;0.13)$$
$$f^4 = (0.48;0.16;0.17;0.16)$$
$$f^5 = (0.50;0.17;0.19;0.18)$$

Таблица 2. Представление набора предложений в евклидовом пространстве

№	МАМА (С)	МЫТЬ (Г)	ОКТЯБРЬ (С)	ПЕРВОЕ (Ч)	РАМА (С)	ТРЯПКА (С)
0	0,33	0,33	0,00	0,00	0,33	0,00
1	0,20	0,20	0,48	0,48	0,20	0,00
2	0,33	0,33	0,00	0,00	0,33	0,00
3	0,25	0,25	0,00	0,00	0,25	0,60

Можно также предположить, что связи между предложениями одного документа, а также связи между предложениями различных документов набора должны быть дифференцированы. В этом случае полагается, что:

$$W = W_{inner} + W_{\text{int}ra}, \quad (5)$$

где W_{inner} – матрица весов связей предложений внутри одного документа, а $W_{\text{int}ra}$ – матрица весов связей предложений разных документов.

В этом случае целесообразно ввести коэффициенты:

$$W = \lambda_1 \cdot W_{inner} + \lambda_2 \cdot W_{\text{int}ra}. \quad (6)$$

2.3 Алгоритм усечения сходных предложений

Для формирования итогового обзорного реферата по набору документов необходимо выполнить следующее:

исключить из рассмотрения предложения, повторяющие по своей структуре те, что уже попали в обзорный реферат.

Выполнить итоговую сортировку предложений с целью получения более-менее связного текста.

Для задачи внесения в итоговый ранг фактора «информационной новизны» используется следующий алгоритм:

[1] Инициализируются два множества $A = \varnothing$ и $B = \{x_i \mid i = 1,2,..n\}$. Для каждого предложения В текущий ранг принимается равным f_i^*.

$$RankScore(x_i) = f_i^*, i = 1,2,..n \quad (7)$$

[2] Предложения множества В сортируются в соответствии с их текущим рангом в порядке убывания.

[3] Полагая, что предложение $\overline{x_i}$ имеет наивысший ранг, оно перемещается из В в А. Ранг оставшихся в В предложений $\overline{x_j}$ рассчитывается как

$$RankScore(x_j) = RankScore(x_j) - \omega \cdot \overline{S}_{j,i} \cdot f_i^* \quad (8)$$

Где $\omega > 0$ – фактор усечения сходных предложений, а

$$\overline{S} = D^{-1} \cdot W \quad (9)$$

[4] Процесс повторяется, пока В не станет пустым.

Вопрос окончательной сортировки предложений в обзорном реферате является темой отдельного исследования. Следует отметить, что в данном случае возможны как лексико-ориентированные алгоритмы, основанные на более детальном анализе предложений, так и семантические алгоритмы, основанные на сравнении поверхностно-семантических графов. Последние также позволяют исключать лексически разные, но семантически подобные предложения, однако не обладают такой простотой, быстротой и универсальностью, как первые. Строго говоря, темой отдельного исследования может быть вектор параметров, влияющих на окончательный порядок следования предложений в итоговом реферате, таких как количество слов в предложении, количество значимых слов, принадлежность предложения к определенному документу, дата опубликования документа, относительная сложность восприятия предложения и т.д.

3 Реализация

Система автоматического реферирования новостных сюжетов реализована как набор скриптов на языке PHP с Web-интерфейсом. Для матричных вычислений было разработано специальное расширение для языка PHP *php_math* на основе MTL[8], позволяющее осуществлять быстрые матричные вычисления. Для морфологического анализа была задействована библиотека phpmorphy, основанная на словарях проекта AOT[7]. Система позволяет подбирать параметры $\alpha, \lambda_1, \lambda_2, \omega$ и количество предложений, попадающих в обзорный реферат. По исходному корпусу строится обзорный реферат с выводом всех промежуточных расчетов. Система позволяет получать обзорные рефераты с заданными параметрами на лету. Система доступна по адресу http://openthesaurus.ru/manifold/.

4 Исходные данные

В качестве исходных данных для оценки работы алгоритма был взят набор кластеров новостной тематики, любезно предоставленный НИВЦ МГУ.

Для кластера «На севере Омской области выпал разноцветный снег» содержащего 8 документов (всего 61 предложение) был получен обзорный реферат из 4 предложений при значении параметров $\alpha = 0.6, \lambda_1 = 0.3, \lambda_2 = 1, \omega = 15$.

«*Представители властей заявили, что если вдруг выяснится, что разноцветный снег в Сиби-*

ри выпал из-за промышленных выбросов, нарушителей привлекут к уголовной ответственности. Пока специалисты только говорят, что аномальные осадки не опасны для здоровья. Кроме того, необычный снег выпал в Томской и Тюменской областях. Вчера были обнародованы первые лабораторные исследования выпавшего 31 января в Омской области желто-оранжевого снега».

5 Оценка

Предварительная оценка работы алгоритма позволяет утверждать о возможности применения его в модифицированном виде для корпусов новостных сюжетов на русском языке. На сегодняшний день на основе ручных аннотаций, любезно предоставленных НИВЦ МГУ (Б.В. Добров) проведена оценка качества системы реферирования при помощи меры ROUGE.

$$ROUGE-N=\frac{\sum_{S\in \mathrm{Re}\, fSum}\ \sum_{n-gram\in S} Count_{match}(n-gramm)}{\sum_{S\in \mathrm{Re}\, fSum}\ \sum_{n-gram\in S} Count(n-gramm)} \quad (10)$$

Мера ROUGE-N представляет собой обобщенную статистическую меру, выражающую какой процент лексических единиц (N-gram,- последовательностей из N лексем), входящих в состав ручной, построенной независимым экспертом, аннотации, попадает в обзорный реферат.

5.1 Результаты оценки

Для предварительной оценки качества работы системы были вычислены меры ROUGE-1, ROUGE-2, ROUGE-3 нескольких обзорных рефератов, полученных при помощи системы, для которых есть ручная аннотация. Для получения рефератов были использованы следующие значения параметров: $\alpha = 0.6, \lambda_1 = 0.3, \lambda_2 = 1, \omega = 15$.

Результаты оценки представлены в таблице 3. Полученные результаты можно сравнить с результатами, полученными в DUC [5] (таблица 4).

Таблица 3. Результаты оценки

№	Тема кластера	ROUGE-1	ROUGE-2	ROUGE-3
1	В Гальском районе Абхазии неизвестные похитили главу районной избирательной комиссии	0.4286	0.2545	0.2037
2	Китай успешно запустил на орбиту навигационный спутник "Бэйдоу"	0.2581	0.0656	0.0333
3	Секретаршу из Coca-Cola признали виновной в краже секретов компании	0.5600	0.4286	0.3542
4	На севере Омской области выпал разноцветный снег	0.4655	0.3157	0.2500

Таблица 4. Сравнение с DUC

	DUC 2003	DUC 2005	Построенная система
ROUGE-1	0.37332	0.38434	**0.42805**
ROUGE-2	0.07677	0.07317	**0.26610**

Приведенные оценки являются предварительными и требуют дальнейшей корректировки с учетом большего количества обзорных рефератов.

6 Будущая работа

Для улучшения качества работы системы необходимо выполнить следующие действия:

[1] Улучшить алгоритм распознавания и разрешения анафор.

[2] Добавить в систему синонимию. Например, лексемы «км» и «километр», «15» и «пятнадцать» должны рассматриваться как совершенно идентичные. Для других вариантов возможно введения коэффициента синонимии.

[3] Провести более основательную оценку качества работы системы на основе большего количества ручных рефератов.

7 Заключение

Задача автоматизации реферирования текстовой информации на сегодняшний день остается очень актуальной. Алгоритм ранжирования связных структур зарекомендовал себя с положительной стороны как довольно эффективный для задач автоматического аннотирования, и в то же время, относительно легко реализуемый, и может применяться для автоматического реферирования корпусов новостных сюжетов на русском языке. Однако алгоритм требует дополнительных доработок с учетом особенностей как русского языка,

так и естественного языка вообще. Так, необходимо более детально рассматривать предложения, содержащие прямую речь, разработать более совершенный (чем в [3]) алгоритм разрешения анафор, а также обеспечить связность текста полученного обзорного реферата.

Литература

[1] Hahn U., Mani I. "The Challenges of Automatic Summarization," Computer, vol.33, no.11, pp. 29-36, Nov., 2000. http://doi.ieeecomputersociety.org/10.1109/2.881692

[2] Белоногов Г.Г., Калинин Ю.П., Хорошилов А.А. Компьютерная лингвистика и перспективные информационные технологии. – М. : Русский мир, 2004. – 246 с.

[3] Н.Н. Абрамова, В.Е. Абрамов. Автоматическое составление обзорных рефератов новостных сюжетов // Труды 9-й Всероссийской научной конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007, Переславль-Залесский, Россия, 2007.

[4] Zhou et al., 2003b D. Zhou, J. Weston, A. Gretton, O. Bousquet and B. Schlkopf. Ranking on data manifolds. In Proceedings of NIPS'2003.

[5] "Manifold-Ranking Based Topic-Focused Multi-Document Summarization" DUC 2003 http://www.ijcai.org/papers07/Papers/IJCAI07-467.pdf

[6] Barzilay R. Sentence Ordering in Multidocument Summarization. Computer Science at Columbia University, Web seit, 2007, http://www.cs.columbia.edu/nlp/papers/2001/barzilay_al_01.pdf

[7] Автоматическая обработка текста. http://aot.ru/

[8] The Matrix Template Library. http://www.osl.iu.edu/research/mtl/

[9] Luhn The Automatic Creation of Literature Abstracts (context) – 1958.

Automatic compilation of news stories reviews

S.D. Tarasov

This work deal with one of the topical problems of automatic summarization – multi-document summarization in respect to news stories. This paper presents a novel extractive approach based on manifold-ranking of sentences to this summarization task. The manifold-ranking algorithm differentiates the intra-document and inter-document links between sentences with different weights. The possibility of the use the algorithm for Russian language is analyzed. A sample system for automatic summarization is build. This paper represents the sample summaries and describes experiments of summarization evaluation. The main problems of implementation of the system and possible methods of their solutions are formulated. The ROUGE criteria was used for evaluation. The results of work of built system are compared with the results of DUC 2003, DUC 2005.

♣ Автор выражает благодарность Б.В. Доброву (НИВЦ МГУ) за предоставленные материалы ручных аннотаций кластеров, а также за помощь в написании данной статьи.