МНОГОАСПЕКТНЫЙ ДАЙДЖЕСТ И СПОСОБЫ ЕГО ФОРМИРОВАНИЯ

Зубова И.И.

Минский государственный лингвистический университет

Аннотация: В статье рассматривается суть понятий и основные характеристики простого и веерного многоаспектного дайджеста. Описывается общая процедура формирования компьютером обычного обзорного реферата. Отмечаются наиболее популярные подходы к автоматическому построению многоаспектного реферата.

Ключевые слова: вес, многоаспектный дайджест, кластер, простой дайджест, текст, тема.

В рамках современных технологий *Text Mining* активно развивается автоматическое реферирование, главная задача которого сводится к извлечению компьютером наиболее важных сведений из большого массива первичных документов и формированию на их основе краткого и информационно емкого обзорного реферата или дайджеста (вторичного документа). Простой дайджест позволяет получить сжатое представление содержания группы текстов одной тематики, в котором должны быть все основные вопросы, затрагиваемые в каждом тексте, но в обобщенном виде, без повторения информации. Путем объединения рефератов, составленных по каждому тексту первичного документа, не удается достичь хорошего результата, так как в таком большом количестве вторичных документов будет содержаться избыточная информация. Поэтому целесообразно выбрать из информационного потока наиболее весомые документы и на их основе сформировать дайджест [2, с. 201]. Текст автоматически созданного дайджеста должен соответствовать определенным критериям. К ним относится четкая структура обзорного реферата и разбиение его на абзацы. Переход от более общих аспектов к более конкретным проблемам

затронутой темы должен быть постепенным. Необходимо, чтобы текст дайджеста был читабелен, т.е. в нем не должно быть «информационного шума». Кроме того, исключается наличие ссылок на то, что не было упомянуто или объяснено, разрывов текста в предложениях и семантическая избыточность [2, с. 202].

В общем плане построение простого дайджеста состоит из нескольких этапов [1; 2; 3]. На первом этапе отбираются лексические единицы с наибольшим весом, которые входят в массив исходных документов. Как отмечалось выше, из входного потока для формирования дайджеста выбираются наиболее весомые исходные документы. Вес каждого документа определяется с учетом нормированной по длине документа суммы весов отдельных слов, входящих в этот документ. Этап выбора весомого документа состоит из следующих шагов. Сначала определяется вес каждого документа, далее входной поток сортируется по весам. Затем определяются смысловые дубли документов по статистическим критериям. При этом для отдельных документов определяются цепочки ключевых слов и частоты их использования, после чего все цепочки исходных документов сравниваются между собой. На следующем шаге отбрасываются непригодные для построения дайджеста документы и смысловые дубли. Далее из отсортированного и отфильтрованного массива выбирается заранее заданное количество самых весомых документов. На последнем этапе построения дайджеста из отобранных документов выделяются наиболее значимые предложения, и из них составляется единый текст, разделенный на подразделы. Если простой дайджест формируется на основе постоянно обновляющейся информации из сети Интернет, то составляется гипертекстовое представление самого дайджеста, который можно считать самостоятельным документом со ссылками на первоисточники в Сети.

Существует такая разновидность простого дайджеста как веерный многоаспектный дайджест, отражающий не только основные аспекты

(темы) входного информационного потока, но и некоторые дополнительные аспекты, не принимающиеся во внимание при создании простого дайджеста [2, с. 203]. При формировании веерного многоаспектного дайджеста применяются те же методы, что и при построении обычного дайджеста. Так, на первом этапе по описанным выше правилам строится простой дайджест, отражающий главную тему всего массива документов. На втором этапе из входного потока текстов удаляются те, которые передают определенный на предыдущем этапе главный аспект. На третьем этапе строится обычный дайджест, но отражающий главный аспект оставшейся части документов. На четвертом этапе полученные простые дайджесты объединяются. Далее выполняется переход ко второму этапу, если в результате требуется дайджест большего объема [2, с. 204–205].

Некоторые наиболее популярные подходы к автоматическому формированию простого и многоаспектного дайджеста описаны в работах [3; 4; 5]. Так, функциональный метод предполагает экстрагирование наиболее релевантных предложений из текста первичного документа и составление из них дайджеста с учетом следующих факторов:

1. частоты слов – важные слова встречаются в тексте многократно;
2. учета слов из заголовка – появление в предложении слова из заголовка указывает на то, что в смысловом плане предложение очень важно для этого текста;
3. позиции предложения – первые предложения текста, как правило, содержат наиболее важную информацию;
4. длины предложения – слишком короткие предложения обычно не включаются в дайджест, так как содержат мало информации; слишком длинные предложения также не подходят для текста вторичного документа;
5. наличия ключевых слов – в предложении могут встречаться определенные слова, указывающие на то, что оно содержит важную информацию, например, *следовательно, в заключении*;

6. наличия имен собственных – предложения, содержащие имена собственные, называющие уникальный объект, например, имя человека, организации или места, в информационном плане считаются важными.

Для формирования простого или многоаспектного дайджеста довольно часто используется кластерный метод [4; 5]. Идея кластеризации заключается в группировании похожих объектов по классам. При формировании дайджеста такими объектами являются предложения, а классами – кластеры, к которым принадлежат эти предложения. Очень похожие друг на друга предложения группируются в один кластер. После этого для формирования окончательного варианта обычного или многоаспектного дайджеста из каждого кластера выбирается одно предложение. Алгоритмы кластеризации бывают агломерационными и разделительными [4]. При агломерационной кластеризации каждое предложение изначально считается отдельным самостоятельным кластером. Далее отдельные кластеры объединяются в большие группы. Этот процесс повторяется до тех пор, пока не будет достигнут какой-либо заранее заданный критерий. При разделительной кластеризации изначально все предложения считаются одним большим кластером, который затем делится на несколько более мелких подкластеров. Суть следующего метода формирования простого или многоаспектного дайджеста заключается в построении графа, отражающего связи между предложениями, исходя из их сходства [4]. После того, как для группы первичных документов сформирован граф, компьютер может определить важные в семантическом плане предложения. Предложение считается информационно значимым, если оно связано со многими другими предложениями.

С целью улучшения качества автоматически создаваемого дайджеста иногда используется метод, опирающийся на знания и учитывающий скрытую семантическую информацию [4; 5]. Однако, несмотря на определенные преимущества, этот подход сложен в реализации, поскольку

в отличие от предыдущих методов, которые могут быть применены к текстам любой тематики, данный подход требует разработки отдельной базы данных для каждой предметной области.

ЛИТЕРАТУРА

1. Абрамова, Н. Н. Автоматическое составление обзорных рефератов новостных сюжетов / Н. Н. Абрамова, В. Е. Абрамов // Труды 9-ой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL’2007. – Переславль-Залесский, 2007. – С. 11–16.
2. Ландэ, Д. В. Поиск знаний в Internet. Профессиональная работа : пер. с англ. / Д. В. Ландэ. – М. : Издательский дом «Вильямс», 2005. – 272 с.
3. Лукашевич, Н. В. Автоматическое аннотирование новостных кластеров на основе тематического представления / Н. В. Лукашевич, Б. В. Добров [Электронный ресурс]. – Режим доступа : http://www.dialog-21.ru/digests/dialog2009/materials/html/46.htm. – Дата доступа : 28.04.2018.
4. Kumar, Y. Automatic Multi Document Summarization Approaches / Y. Kumar, N. Salim // Computer Science [Electronic resource]. – Mode of access: https://pdfs.semanticscholar.org/60d1/b62b206144bb786f79baa04e1702ef0daeb1.pdf. – Date of access : 05.05.2018.
5. Mani, I. Automatic Summarization / I. Mani // Natural Language Processing (Ed. R. Mitkov). – Philadelphia/Amsterdam, 2001. – Vol. 3. – 286 p.