Московский государственный университет имени М.В. Ломоносова

На правах рукописи

Алексеев Алексей Александрович

Метод автоматического аннотирования новостных кластеров на основе тематического анализа

Специальность 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Автореферат

диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2014

Работа выполнена на кафедре алгоритмических языков факультета вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова.

Научный руководитель: доктор физико-математических наук, профессор, зав. каф. алгоритмических языков ВМиК МГУ имени М.В. Ломоносова, Мальковский Михаил Георгиевич

Официальные оппоненты: Фомичев Владимир Александрович
доктор технических наук, профессор,
НИУ ВШЭ, факультет бизнес-информатики,
профессор кафедры инноваций и бизнеса в сфере информационных технологий

Васильев Виталий Геннадьевич
кандидат технических наук, доцент,
ООО «ЛАН-ПРОЕКТ», научный консультант

Ведущая организация: Казанский (Приволжский) Федеральный Университет

Защита состоится 19 сентября 2014 г. в 11 часов на заседании диссертационного совета Д.501.001.44 при Московском государственном университете имени М.В. Ломоносова по адресу: 119991, ГСП-1, Москва, Ленинские горы, МГУ, 2-й учебный корпус, факультет Вычислительной математики и кибернетики, аудитория 685. Желающие присутствовать на заседании диссертационного совета должны сообщить об этом за 2 дня до указанной даты по тел. (495) 939-30-10 (для оформления заявки на пропуск).

С диссертацией можно ознакомиться в Фундаментальной библиотеке МГУ имени М.В. Ломоносова. С текстом автореферата можно ознакомиться на официальном сайте ВМК МГУ имени М.В. Ломоносова http://www.cmc.msu.ru в разделе «Наука» – «Работа диссертационных советов» – «Д 501.001.44».

Автореферат разослан «__» августа 2014 года.

Ученый секретарь диссертационного
совета Д 501.001.44, к. т. н., в. н. с. Костенко В.А.

Общая характеристика работы

Актуальность темы. Развитие информационных технологий и появление сети Интернет явились причиной экспоненциального роста объемов электронной информации, начавшегося приблизительно два десятилетия назад и стремительно продолжающегося в настоящее время. Объемы информации уже сейчас достигли таких размеров, что человек не способен самостоятельно ознакомиться с материалами всех информационных источников, зачастую даже в контексте специализированных информационных потребностей. Данный факт обусловил активное развитие исследований в области задачи автоматического аннотирования – представления релевантной и наиболее значимой информации, необходимой пользователю, в сжатом, лаконичном виде.

Методы автоматического аннотирования исследовались в трудах российских и зарубежных ученых, таких как Барзилай Р., Добров Б.В., Лукашевич Н.В., Лун Х., МакКьюин К., Мальковский М.Г., Мани И., Машечкин И.В., Ненкова А., Петровский М.И., Севбо И.П., Тарасов С.Д., Фомичев В.А., Шиффман Б., Эдмундсон Х. и многих других авторов. Спектр областей применения систем автоматического аннотирования обширен, от бытовых информационных потребностей обычных пользователей до узкоспециализированных аналитических задач. Например, в рамках исследовательской программы SUMMAC[1] (США) установлено, что время принятия аналитиком решения о релевантности текстового документа некоторой тематике может быть сокращено в 2 раза за счет использования аннотации исходного документа, без статистически значимого ухудшения точности данного решения. Подготовка обзорных рефератов для коллекции документов уже давно является одним из ключевых элементов в организации

[1] http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac/

и представлении результатов поиска, основным показателем эффективности которого является минимизация его общего времени.

При этом как сами задачи аннотирования, так и предметные области бывают достаточно специфичны, что зачастую влечет за собой необходимость разработки индивидуальных алгоритмов аннотирования. Современные технологии автоматической обработки новостных потоков основаны на тематической кластеризации новостных сообщений, т. е. выделении совокупностей новостей, посвященных одному и тому же событию – ***новостных кластеров***. Одной из важных и актуальных специализированных задач аннотирования является ***автоматическое аннотирование новостных кластеров***. Новостной кластер и методы автоматического аннотирования новостных кластеров являются основными объектами рассмотрения данной кандидатской диссертации.

Новостной кластер должен соответствовать ситуации (или совокупности связанных ситуаций), со своим набором участников, которые в исходном кластере могут быть представлены различными ***языковыми выражениями***, то есть словами или словосочетаниями. Например, международный аэропорт «Внуково», расположенный в Москве, может упоминаться в рамках некоторого новостного кластера как *московский международный аэропорт Внуково*, *московский аэропорт*, *столичный аэропорт*, *аэропорт Внуково*, *международный аэропорт* и так далее.

Качественное выделение участников ситуации, включая различные варианты их наименования в различных документах кластера, может помочь лучше определять основную тему новостного кластера и, таким образом, позволит повысить качество различных операций с новостными кластерами, таких как автоматическое аннотирование, определение новизны информации и других автоматических операций.

Таким образом, актуальной является задача выявления различных вариантов именования основных участников ситуации, описываемой в рамках новостного кластера. В данной работе предлагается модель

представления содержания новостного кластера, описывающая основных участников ситуации с учетом вариативности их именования – тематических цепочек новостного кластера. Рассматриваются методы улучшения качества извлечения основных участников новостного события, что включает нахождение совокупности слов и выражений, с помощью которых тот или иной значимый участник события именовался в документах новостного кластера. Предлагаемый подход основан на совместном использовании совокупности факторов, в том числе разного рода контекстов употребления слов в документах кластера, информации из предопределенных источников (тезаурус русского языка), а также особенностях построения текстов на естественном языке.

Целью диссертационной работы являются разработка методов и программных средств построения модели основных участников новостного кластера с учетом вариативности их именования на основе комбинации разнородных факторов схожести и интеграция построенной модели в методы автоматического аннотирования. Разрабатываемые программные средства и полученная модель должны удовлетворять следующим требованиям: высокая точность выявления различных вариантов именования основных участников; возможность интеграции построенной модели в другие задачи автоматической обработки текста; независимость от предметной области.

Для достижения этой цели были решены следующие **задачи**:

1. исследование и построение модели основных участников новостного кластера с учетом вариативности их именования и специфики внутреннего устройства текстов на естественном языке;
2. разработка методов интеграции построенной модели в методы автоматического аннотирования, а также разработка двух новых метод на основе построенной модели;
3. разработка и реализация программного модуля для построения тематических цепочек новостного кластера;

4. разработка и реализация программного модуля автоматического аннотирования новостного кластера, реализующего методы аннотирования на базе построенных тематических цепочек.

Основные положения, выносимые на защиту:

1. Предложен и реализован новый метод автоматического построения модели основных участников новостного кластера (выражаемых тематическими цепочками), основанный на комбинировании разнородных признаков сходства;
2. Предложен метод применения построенной модели в существующих методах автоматического аннотирования;
3. На основе построенной модели предложены и реализованы два новых метода автоматического аннотирования;
4. Показано улучшение качества работы алгоритмов аннотирования на основе тематических цепочек.

Научная новизна настоящей диссертационной работы заключается в том, что предложен новый метод построения модели совокупности участников новостного кластера, основанный на комбинации признаков различной природы: как статистических контекстных, так и априорных. Применимость данного метода обоснована теоретически – на основе анализа полезности отдельных признаков для определения близости языковых выражений, а также численно – на основе экспериментов по интеграции в методы автоматического аннотирования. Разработанная модель не зависит от предметной области и может применяться в различных задачах автоматической обработки новостных кластеров.

Практическая значимость. На основе предложенного алгоритма спроектирована и реализована многомодульная программная система со следующими функциональными возможностями:

- построение тематических цепочек новостного кластера;

- автоматическое формирование аннотаций новостного кластера различными алгоритмами аннотирования;
- автоматическая оценка конкурсных аннотаций.

Таким образом, разработанная система может быть использована как для подготовки дополнительной входной информации для других систем автоматической обработки новостных кластеров, так и для формирования автоматических аннотаций новостного кластера.

Апробация работы. Основные результаты работы докладывались на следующих конференциях и семинарах:

- всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (Казань, 13-17 октября 2010 г.);
- международной конференции «Математика. Компьютер. Образование» (Дубна, 25-30 января 2010 г.);
- семинаре по поиску концептов в неструктурированной информации (CDUD), проходящему совместно с конференцией RSFDGrC (Москва, 25-30 июня 2011 г.);
- международной конференции «Системный анализ и семиотическое моделирование» (Казань, 24-27 февраля 2011 г.);
- международной конференции «Диалог» (Московская область, 25-29 мая 2011 г.);
- летней школе по информационному поиску RUSSIR (Ярославль, 6-10 августа 2012 г.);
- международной конференции «Spring Researchers Colloquium on Databases and Information Systems» (Москва, 1 июня 2012 г.);
- всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (Ярославль, 14-17 октября 2013 г.);

Кроме того результаты обсуждались на семинаре лаборатории анализа информационных ресурсов НИВЦ МГУ, на семинаре в НИУ ВШЭ и на регулярном семинаре ACM SIGMOD в Москве.

Личный вклад автора заключается в выполнении основного объема теоретических и экспериментальных исследований, изложенных в диссертационной работе, включая разработку теоретической модели, методик экспериментальных исследований, проведение исследований, анализ и оформление результатов в виде публикаций и научных докладов.

Основные положения, выносимые на защиту, описанные в совместных публикациях, принадлежат автору диссертации.

Публикации. Основные результаты по теме диссертации изложены в 14 печатных работах, в том числе 3 статьях в журналах из списка ВАК ([1], [2], [3]) и 3 статьях, входящих в базу SCOPUS ([4], [5], [6]).

Объем и структура диссертации. Диссертация состоит из введения, четырех глав, заключения и двух приложений. Полный объем диссертации составляет 122 страницы с 15 рисунками и 7 таблицами, объем приложений – 9 страниц. Список литературы содержит 82 наименования.

Содержание работы

Во **введении** обоснована актуальность диссертационной работы, сформулирована цель исследований, показаны их научная новизна и практическая значимость, представлены научные положения, выносимые на защиту.

Первая глава посвящена описанию задачи автоматического аннотирования, классификации типов аннотаций и областей применимости систем автоматического аннотирования. Также в данной главе приводится обзор алгоритмов построения аннотаций, базовых идей, моделей и принципов построения автоматических аннотаций, а также методов оценки качества и сравнения результатов работы различных систем автоматического аннотирования. Особое внимание уделяется задаче и алгоритмам построения

обзорных рефератов, так как основным объектом исследования данной кандидатской диссертации является новостной кластер. Целью данной главы является анализ достоинств и недостатков существующих методов автоматического аннотирования, а также проблем в данной области, в частности обоснование важности учета лексическо-семантической вариативности, широко присутствующей в текстах на естественном языке.

Подавляющее число современных систем автоматического аннотирования работает на основе экстрактивного подхода к аннотированию, т. е. выбора целых предложений исходной коллекции S для автоматической аннотации S`, удовлетворяющей ограничению на длину L:

$$S` = \arg\max_{S^* \subset S} f(S^*) \wedge Length(S^*) < L$$

где S^* - подмножество предложений S, $f()$ – некоторая мера качества для аннотации. Оценка информативности предложений, на базе которой происходят ранжирование и отбор предложений в результирующую аннотацию, строится на основе слов и выражений, входящих в данное предложение. При этом большинство методов автоматического аннотирования работает с пословной моделью представления предложений входной коллекции:

$$S = \{s = [w_1, .., w_m] \mid m \geq 1\},$$

где $w \in W$ - словарь, m – количество слов в предложении $s \in S$.

Данная модель обладает рядом ограничений, связанных с лексико-семантической вариативностью, широко встречающейся в текстах на естественном языке. Это означает, что для качественного решения задачи автоматического аннотирования необходима более комплексная модель, содержащая информацию о:

- языковых выражениях $T = \{t_i = [w_{i1}, .., w_{in}] \mid n \geq 1, i = 1...M\}$ - словах и многословных выражениях, где n – количество слов содержащихся в языковом выражении t_i, M – общее количество языковых выражений в рассматриваемой коллекции;

- соответствии языковых выражений *T* сущностям исходной коллекции.

<u>Вторая глава</u> посвящена проблеме вариативности терминов в текстах на естественном языке, описываются существующие методы выявления различных типов вариативности, а также вводится формальная модель совокупности участников ситуации, описываемой в текстовой коллекции, с учетом вариативности их упоминаний – ***тематических цепочек***. Предлагаемые модель и алгоритм построения тематических цепочек основаны на свойствах глобальной и локальной связности текстов на естественном языке. Ван Дейк и Кинч[2] описывают тематическую структуру текста как иерархическую, в том смысле что тема всего текста описывается посредством более конкретных подтем, которые в свою очередь могут быть охарактеризованы посредством еще более конкретных подтем текста и т.д. Под темой/подтемой при этом понимается предикат $P(\mathrm{C}_1,\ldots,\mathrm{C}_\mathrm{n})$, его атрибуты $C_1,\ldots,C_n$ будем называть ***тематическими элементами***. Каждое предложение *s* связного текста посвящено раскрытию той или иной подтемы P_i уровня *level* основной темы текста: $s \to \mathrm{P}_i^{level}(\mathrm{C}_{\mathrm{i\_1}}^{\mathrm{level}},\ldots,\mathrm{C}_{\mathrm{i\_n}}^{\mathrm{level}})$, раскрывающей один из аспектов взаимоотношений тематических элементов $\mathrm{C}_{\mathrm{i\_1}}^{\mathrm{level}},\ldots,\mathrm{C}_{\mathrm{i\_n}}^{\mathrm{level}}$. При этом отнесение к тематическим элементам $\mathrm{C}_{\mathrm{i\_1}}^{\mathrm{level}},\ldots,\mathrm{C}_{\mathrm{i\_n}}^{\mathrm{level}}$ внутри *s* осуществляется с помощью конкретных языковых выражений, упомянутых в *s*:

$$s \leftrightarrow \mathrm{P}_i^{level}(\mathrm{C}_{\mathrm{i\_1}}^{\mathrm{level}},\ldots,\mathrm{C}_{\mathrm{i\_n}}^{\mathrm{level}}) = P_i^{level}(\mathrm{C}_{\mathrm{i\_1}}^{\mathrm{level}} \to \{\mathrm{t}_{\mathrm{i\_1}}^{\mathrm{level}}\},\ldots,\mathrm{C}_{\mathrm{i\_n}}^{\mathrm{level}} \to \{\mathrm{t}_{\mathrm{i\_n}}^{\mathrm{level}}\}), \quad \mathrm{t}_{\mathrm{i\_1}}^{\mathrm{level}},\ldots,\mathrm{t}_{\mathrm{i\_n}}^{\mathrm{level}} \in s$$

Таким образом, для отнесения к некоторому тематическому элементу C_m используется определенный набор языковых выражений, каждое из которых применяется для раскрытия определенных подтем текста: $\mathrm{C}_\mathrm{m} \to \{\mathrm{t}_\mathrm{m}^1,\ldots,\mathrm{t}_\mathrm{m}^\mathrm{i},\ldots\}$.

[2] Дейк В., Кинч В. Стратегии понимания связного текста // Новое в зарубежной лингвистике, Выпуск 23, Москва, 1988. - С. 153-211.

Из описанной модели следует важный практический вывод: если языковые выражения t_1 и t_2 часто встречаются в анализируемом тексте в одних и тех же простых предложениях, то это означает, что данный текст посвящен рассмотрению отношений между этими сущностями, т. е. t_1 и t_2 соответствуют разным тематическим элементам. С другой стороны, если два языковых выражения t_1 и t_2 редко встречаются в одних и тех же предложениях текстов, но при этом часто упоминаются в соседних предложениях, то это дает возможность предположить, что они используются для осуществления локальной связности, то есть между ними имеется смысловая связь.

Гипотеза о совместной встречаемости связанных языковых выражений легла в основу ограничивающего фактора $IsNSCriterion(t_i,t_j)$, который является управляющим в предлагаемом алгоритме построения тематических цепочек:

$$count((s_k, s_m) \mid t_i \in s_k \wedge t_j \in s_m \wedge NS(s_k, s_m)) > C \cdot count(s \mid t_i \in s \wedge t_j \in s)$$

где $count(A|B)$ – количество элементов A, удовлетворяющих условию B (в данном случае предложений и пар предложений); $NS(s_k, s_m)$ – признак последовательного появления предложений s_k и s_m в исходном новостном кластере. Необходимо отметить что критерии подобные критерию $IsNSCriterion(t_i,t_j)$, не использовалась ранее для решения таких задач, как установление вариантов именования основных участников ситуации, построение рядов квазисинонимов, лексических цепочек и т.п. Таким образом, задача построения тематических цепочек представляет собой задачу ***кластеризации с ограничениями***:

- $TC = \{tc_i \in a^M\}, a_{ij} \in \{0, 1, 2\}, M = |T|$, где tc_i – тематическая цепочка языковых выражений с выделенным центральным элементом (кластер языковых выражений);

Ограничения:

- $\forall i : count_{j=1..M}([tc_{ij} = 2]) = 1$ – каждая тематическая цепочка содержит один и только один центральный элемент;

- $\forall j : (sum_1, .., sum_M) = \sum_{tc_i \in TC} tc_{ij} : 1 \le sum_j \le 2$ – каждое языковое выражение является элементом не более чем двух и не менее чем одной тематической цепочки либо центром единственной тематической цепочки;
- $\forall tc_i : (tc_{ij} > 0 \wedge tc_{ik} > 0) \Rightarrow \mathrm{IsNSCriterion}(tc_{ij}, tc_{ik}) = true$ – выполнено ограничивающее условие на объединение языковых выражений в тематическую цепочку.

В диссертации предложен алгоритм построения тематических цепочек, объединяющий характеристики схожести различной природы – контекстно-зависимых и контекстно-независимых признаков. Каждая характеристика имеет вещественный вес в диапазоне [0,1].

Контекстно-зависимые характеристики

Количество вхождений в соседние предложения (**N**eighboring **S**entence **F**eature, ***NSF***). Данная характеристика основана на гипотезе глобальной связности текстов на естественном языке и её следствии о том, что элементы одной тематической цепочки чаще появляются в соседних предложениях исходных документов, чем в одних и тех же предложениях.

Характеристика *NSF* вычисляется на основе контекстных параметров *AcrossVerb* (количество вхождений в одно предложение через глагол), *Near* (количество вхождений в одно предложение непосредственно рядом), *NotNear* (количество вхождений в одно предложение не рядом) и *NS* (количество вхождений в соседние предложения), а также распределения их средних значений внутри исходного новостного кластера:

$$C(t_i, t_j) = NS(t_i, t_j) - 2 \cdot \left(AcrossVerb(t_i, t_j) + Near(t_i, t_j) + NotNear(t_i, t_j) \right)$$

$$\underset{t_i, t_j \in T}{weight}_{NSF}(t_i, t_j) = \min(1, \frac{C(t_i, t_j)}{\underset{t_k, t_m \in T}{Avg}(C(t_k, t_m))})$$

где *Avg(C)* является средним значением *C* среди всех положительных значений в рамках кластера.

Строгие контексты (**S**trict **C**ontext, ***SC***). Данная характеристика основана на сравнении строгих контекстов употреблений слов – текстовых шаблонов. В качестве шаблонов рассматриваются 4-граммы, два выражения влево и вправо от рассматриваемого выражения: $s_i = (t_{i1}, .., t_{ij-2}, t_{ij-1}, t_{ij}, t_{ij+1}, t_{ij+2}, ..)$, где $(t_{ij-2}, t_{ij-1}, t_{ij+1}, t_{ij+2})$ является строгим контекстом выражения t_{ij} в некотором предложении s_i. Итоговая схожесть по характеристике *SC* для выражений t_i и t_j имеет следующий вид:

$$\underset{t_i, t_j \in T}{weight}_{SC}(t_i, t_j) = \frac{\sum_{templ \in TEMPLATES(t_i) \cap TEMPLATES(t_j)} weight(templ)}{\max\limits_{t_k, t_m \in T} \sum_{templ \in TEMPLATES(t_k) \cap TEMPLATES(t_m)} weight(templ)}$$

Схожесть контекстов употребления по внутренним характеристикам предложения (**S**calar **P**roduct **S**imilarity, ***SPS***). Анализу подвергаются вектора контекстов сравниваемых языковых выражений, сравнение производится по классической косинусной мере:

$$t_i \rightarrow \left| \begin{array}{l} \overrightarrow{V_i^{AcrossVerb}} = (v_{i\_1}^{AcrossVerb}, .., v_{i\_m}^{AcrossVerb}) \\ \overrightarrow{V_i^{Near}} = (v_{i\_1}^{Near}, .., v_{i\_m}^{Near}) \\ \overrightarrow{V_i^{NotNear}} = (v_{i\_1}^{NotNear}, .., v_{i\_m}^{NotNear}) \\ \overrightarrow{V_i^{NS}} = (v_{i\_1}^{NS}, .., v_{i\_m}^{NS}) \end{array} \right. \qquad \underset{t_i, t_j \in T}{weight}_{SPS}(t_i, t_j) = \frac{(\overrightarrow{V_i^{Context}} 1, \overrightarrow{V_j^{Context}})}{|\overrightarrow{V_i^{Context}}| \cdot |\overrightarrow{V_j^{Context}}|}$$

где *Context*={*AcrossVerb*, *Near*, *NotNear*, *NS*} – различные типы контекстов.

Контекстно-независимые характеристики

Формальное сходство (**B**eginning **S**imilarity, ***BS***). Рассмотрение формального сходства выражений является естественным путем обнаружения семантически-связанных объектов. В существующей реализации используется простая мера схожести – одинаковые начала слов. Сопоставление языковых выражений происходит на основе модифицированной меры Жаккара:

$$\underset{t_i,t_j \in T}{weight}_{BS}(t_i,t_j) = \begin{cases} \dfrac{n_{word}(t_i \cap t_j)+k}{n_{word}(t_i \cup t_j)+k} \; при \; n_{word}(t_i \cap t_j) > 0, \; k = 3, \\ 0, \; при \; n_{word}(t_i \cap t_j) = 0. \end{cases}$$

Информация о схожести, описанная во внешнем ресурсе – тезаурусе РуТез (**T**hesaurus **S**imilarity, ***TS***). Анализ информации из внешнего ресурса – тезауруса РуТез, а именно, следующих видов связей: синонимия, часть – целое, род – вид. Вес схожести убывает с ростом длины пути по отношениям и имеет следующий вид:

$$\underset{t_i,t_j \in T}{weight}_{TS}(t_i,t_j) = f_{path}(t_i,t_j) = f(N_{rel}(t_i,t_j),\{\mathrm{Re}l_{type}(t_i,t_j)\})$$

где N_{rel} – длина пути по отношениям тезауруса (количество связей), $\{Rel_{type}\}$ – информация о типах связей по данному пути.

Наличие одинаковых языковых выражений (**E**mbedded **O**bjects **S**imilarity, ***EOS***). При анализе схожести тематических цепочек, включающих в себя несколько языковых выражений, важным фактором схожести является наличие общих языковых выражений:

$$\underset{tc_i,tc_j \in TC}{weight}_{EOS}(tc_i,tc_j) = \begin{cases} 1 \; при \; count(tc_i \cap tc_j) > 0, \\ 0 \; при \; count(tc_i \cap tc_j) = 0. \end{cases}$$

Алгоритм построения тематических цепочек является итеративным, в рамках каждой из итераций происходят ранжирование всех пар – кандидатов на объединение – по суммарному весу характеристик схожести и объединение одной тематической цепочки. Итеративный процесс продолжается до тех пор, пока есть пары – кандидаты для объединения с весом схожести выше установленного порога. Необходимым предварительным этапом построения тематических цепочек является сборка многословных выражений, которая основана на естественном принципе превышения встречаемости слов непосредственно рядом друг с другом по сравнению с раздельной встречаемостью:

$$\mathrm{Near} > 2 \cdot (\mathrm{AcrossVerb} + \mathrm{NotNear}).$$

Например, тематическая цепочка с центральным элементом *пост* проходит следующие этапы в процессе построения:

Итерация 7: (***Отставка)*** ← ***(Отставка с должности)***

Итерация 33: (***Отставка,*** *Отставка с должности)* ← ***(Уход в отставку)***

Итерация 44: (***Отставка,*** *Отставка с должности, Уход в отставку)* ← ***(Отставка президента)***

Итерация 61: (***Уход с поста)*** ← ***(Уход в отставку)***

Итерация 62: (***Отставка,*** *Отставка с должности, Уход в отставку, Отставка президента)* ← ***(Уход с поста,*** *Уход в отставку)*

Итерация 102: (***Отставка,*** *Отставка с должности, Уход в отставку, Отставка президента, Уход с поста)* ← ***(Пост)***

Итерация 103: (***Пост,*** *Отставка, Отставка с должности, Уход в отставку, Отставка президента, Уход с поста)* ← ***(Должность)***

Итерация 104: (***Пост,*** *Отставка, Отставка с должности, Уход в отставку, Отставка президента, Уход с поста, Должность)* ← ***(Уход)***

Псевдокод алгоритма построения тематических цепочек:

Процедура: Построение тематических цепочек	
∇ Вход:	1. Новостной кластер *D* с выделенными языковыми выражениями *T* 2. *Similarity_Score*(tc_1, tc_2) – общий вес по характеристикам схожести для тематических цепочек tc_1 и tc_2 3. *IsNSCriterion*(tc_1, tc_2) – признак выполнения ограничивающего фактора 4. C_1, C_2, C_3 – параметры алгоритма
∇ Выход:	1. Набор тематических цепочек *TC* новостного кластера *D*

```
// Инициализируем множество тематических цепочек отдельными языковыми выражениями
TC = T;
joinFlag = true;
while(joinFlag)
   joinFlag = false;

   // Сформировать пары цепочек, удовлетворяющих ограничению
   Pairs = {(tc_i, tc_j) | IsNSCriterion(tc_i, tc_j)=true, tc_i, tc_j ∈ TC};
```

```
   // Отсортировать пары по убыванию схожести
   Pairs.OrderByDescending(Similarity_Score(tc_i, tc_j) );

   // Выбрать пару для объединения
   { tc_i, tc_j } = Pairs[0];

   // Объединение в случае достаточной схожести
   if ( Similarity_Score(tc_i, tc_j) > C)
      if ( Frequency(tc_i) > Frequency(tc_j) )
         tc_new={t_main=t_main_i, t_i1, … , t_in, t_j1, … , t_jm};
         TC.Remove(tc_i);
      else
         tc_new={t_main=t_main_j, t_i1, … , t_in, t_j1, … , t_jm};
         TC.Remove(tc_j);
      end-if;

      // Произвести расчет характеристик для новой пары tc_new
      CalculateParameters (D, TC, tc_new);

      TC.Add ( tc_new );
      joinFlag = true;
   end-if;
end-while;
```

Сложность предложенного алгоритма построения тематических цепочек новостного кластера с *m* итерациями работы имеет следующий вид:

$$O(n^2) + m \cdot O(2 \cdot n), \quad где\, n - количество\, языковых\, выражений$$

Одной из базовых задач автоматической обработки естественного языка является установление схожести фрагментов текстов, в частности, установление схожести предложений. Соответствующая операция лежит в основе большинства алгоритмов аннотирования, работающих по принципу выделения наиболее значимых предложений исходной коллекции. В рамках диссертационной работы сформулирована и доказана лемма, описывающая влияние операций, выполняемых при построении модели основных участников ситуации, на установление схожести фрагментов текста.

***Лемма*.** Последовательное применение операций добавления многословного выражения $f_{MWE}(w_i^1, w_j^1, \overrightarrow{s_1})$ и установления схожести $f_{TC}(w_r^1, w_m^2, tc, \overrightarrow{s_1}, \overrightarrow{s_2})$ при выполнении условия на установление схожести для одной из частей

многословного выражения (*)[3] приводит к ***увеличению*** косинусной меры схожести между предложениями

$$f_{MWE}(w_i^1, w_j^1, \overrightarrow{s_1}) \wedge (w_i^1 \in s_2) \wedge (w_j^1 \in s_2) \Rightarrow \exists tc`: (w_i^1 w_j^1 \in tc`) \wedge ((w_i^1 \in tc`) \vee (w_j^1 \in tc`)) \quad (*)$$

Эта лемма подтверждает возможность повышения качества методов автоматической обработки текстов за счет интеграции моделей, описывающих основных участников входной текстовой коллекции.

В **<u>третьей главе</u>** описывается алгоритм интеграции построенной модели тематических цепочек в существующие методы автоматического аннотирования **M**aximal **M**arginal **R**elevance[4] (MMR) и Sumbasic[5]. Интеграция заключается в двухступенчатом переходе от пространства отдельных слов (bag-of-words model) к пространству языковых выражений:

<u>Слова</u> → *<u>Объекты</u> (слова + мног.выр.)* → *<u>Тематические цепочки</u>*

I. Замена слов на многословные выражения. Добавление информации о многословных выражениях – переход от <u>слова</u> к <u>объекту</u> (отдельные слова или многословные выражения);

II. Добавление информации о принадлежности объектов тематическим цепочкам. В рамках предлагаемой модели тематических цепочек каждый объект может принадлежать к одной или двум цепочкам.

Каждая тематическая цепочка имеет вес, равный сумме частот его элементов:

$$weight(tc) = \sum_{tc_{elem\_i} \in tc} freq(tc_{elem\_i})$$

Элементы цепочек имеют вес схожести с центральным элементом, равный отношению набранного суммарного балла по характеристикам схожести (при

[3] Добавление многословного выражения $w_i^1 w_j^1$ в предложение s_1 в случае вхождения компонентов данного выражения w_i^1 и w_j^1 в предложение s_2 требует установления дополнительной связи нового выражения $w_i^1 w_j^1$ с одним из его компонентов

[4] Carbonell J., Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries // Proceedings of ACM SIGIR`1998, Australia, pp. 335 – 336

[5] Nenkova, A. and L. Vanderwende. The impact of frequency on summarization // Microsoft Research Technical Report, MSR-TR-2005-101, 2005

построении данной тематической цепочки) к максимально возможному баллу схожести:

$$weight(tc_{elem}) = \frac{similarity(tc_{elem}, tc_{center})}{\max\limits_{tc_{elem\_i} \in tc} (similarity(tc_{elem\_i}, tc_{center}))} \cdot weight(tc)$$

Кроме того, на основе сконструированных тематических цепочек предлагаются два новых метода автоматического аннотирования:

- Отбор предложений на основе участников ситуации (по тематическим цепочкам):

$$s_i \Rightarrow \max\left(\sum_{tc_{new\_j} \in s_i,\ j=1..3}^{desc\ weight(tc_{new})} weight\ (tc_{new\_j}) \right)$$

- Отбор предложений на основе взаимоотношений участников ситуации (по связям тематических цепочек):

$$\mathrm{s_i} \supset \max_{tc_{rel\_new} \in Cluster} \left(weight(tc_{rel\_new}) \right)$$

где $tc_{rel} = \{tc_1, tc_2\}$ – пара тематических цепочек; $weight(tc_{rel})$ – число вхождений пары в одни и те же предложения кластера; tc_{rel_new} – новая пара тематических цепочек, не упомянутая в одних и тех же предложениях, уже отобранных в аннотацию.

Оценка качества полученных автоматических аннотаций, а именно, сравнение модификаций методов с интеграцией и без интеграции построенных тематических цепочек является мерой качества самих тематических цепочек. Для проведения оценки были подготовлены 11 новостных кластеров различной тематики (спорт, политика, происшествия), построенных на основе пословной модели. К каждому кластеру профессиональными лингвистами были подготовлены от 2 до 4 ручных аннотаций. Всего оценке подверглись 10 модификаций методов (классические методы аннотирования 4, 5, 9; классические методы с интеграцией тематических цепочек 1, 7, 10; новые методы аннотирования на основе тематических цепочек 2, 3, 6, 9 с учетом и без учета IDF).

Метод	1	2	L	S	SU	Avg
1. MMR + Groups	**0.62499 (1)**	**0.41633 (1)**	**0.60210 (1)**	**0.35529 (1)**	**0.36649 (1)**	**1.0**
2. OurSummary (Nodes)	0.58652 (2)	0.36154 (2)	0.56450 (2)	0.32113 (2)	0.33203 (2)	2.0
3. OurSummary (Nodes) with IDF	0.58497 (3)	0.33918 (4)	0.55745 (3)	0.30124 (3)	0.31283 (3)	3.2
4. Classic MMR	0.55917 (4)	0.34539 (3)	0.54012 (4)	0.29428 (4)	0.30519 (4)	3.8
5. ThematicLines	0.53416 (5)	0.33364 (5)	0.51238 (5)	0.27130 (5)	0.28243 (5)	5.0
6. OurSummary (Relations)	0.53141 (6)	0.28920 (6)	0.50422 (6)	0.25382 (6)	0.26509 (6)	6.0
7. SumBasic + Groups	0.52255 (7)	0.22881 (9)	0.49300 (8)	0.24356 (7)	0.25525 (7)	7.6
8. SumBasic	0.51847 (8)	0.24735 (8)	0.49786 (7)	0.23064 (8)	0.24257 (8)	7.8
9. OurSummary (Relations) with IDF	0.45494 (9)	0.24856 (7)	0.43768 (9)	0.19419 (10)	0.20492 (10)	9.0
10. MMR with IDF + Groups	0.44475 (10)	0.22238 (10)	0.42318 (10)	0.20627 (9)	0.21648 (9)	9.6

Табл. 1: Результаты оценки методом ROUGE

Процедура оценки состояла из двух этапов. Сначала все модификации методов были оценены автоматическими мерами качества официального пакета ROUGE[6]. В Табл. 1 представлены результаты ROUGE по основным мерам качества (Avg – средняя позиция по всем мерам качеств).

Наиболее значимыми являются следующие результаты:

I. Интеграция построенных тематических цепочек в классические методы автоматического аннотирования MMR и SumBasic улучшает качество исходных методов;

II. Методы аннотирования на основе обогащенной модели тематических цепочек показывают лучшее качество по сравнению с методом ThematicLines, основанным на единственной характеристике схожести.

Для подтверждения результатов оценки методом ROUGE лучшие и наиболее приоритетные модификации методов были дополнительно оценены методом «Пирамиды»[7] (Табл. 2).

[6] Lin C.-Y. ROUGE: a Package for Automatic Evaluation of Summaries // Proceedings of ACL’2004, pp. 74-81

[7] Harnly A., Nenkova A., Passonneau R., Rambow O. Automation of summary evaluation by the pyramid method // Proceedings of RANLP’2005, Bulgaria, 2005

Метод	Score
MMR + Groups	**0.645 (1)**
OurSummary (Nodes)	0.602 (2)
Classic MMR	0.578 (3)
SumBasic + Groups	0.575 (4)
SumBasic	0.567 (5)

Табл. 2: Результаты оценки методом «Пирамиды»

Результаты оценки методом «Пирамиды» подтверждают факты, установленные при оценке методом ROUGE, а именно, улучшение качества методов автоматического аннотирования при интеграции в них построенных тематических цепочек на основе совокупности разнородных факторов.

В рамках проведенного диссертационного исследования разработан программный комплекс по автоматической обработке новостных кластеров, описание которого приведено в **четвертой главе**. Данный комплекс включает в себя следующие независимые модули:

- построения тематических цепочек новостного кластера на основе разработанного алгоритма;
- автоматического аннотирования, реализующий более 10 различных методов аннотирования;
- автоматической оценки аннотаций новостного кластера на основе метода ROUGE.

Модули объединены в единое приложение и могут взаимодействовать друг с другом по принципу конвейера в указанной последовательности, обеспечивая замкнутый цикл обработки новостного кластера всеми функциональными блоками.

В **заключении** приведены основные результаты работы, которые состоят в следующем:

1. Предложена модель, позволяющая с помощью тематических цепочек новостного кластера описывать основных участников этого кластера с учетом вариативности их именования и специфики внутреннего устройства текстов на естественном языке;

2. Предложен и реализован новый метод автоматического построения тематических цепочек новостного кластера, основанный на комбинировании разнородных признаков схожести;
3. Предложен и реализован метод интеграции построенной модели в существующие методы автоматического аннотирования, а также два новых метода автоматического аннотирования на основе тематических цепочек. Показано улучшение качества работы алгоритмов аннотирования на основе построенной модели.

Основные публикации автора по теме диссертации

Издания из списка ВАК:

[1] Алексеев А.А. Тематический анализ новостного кластера как основа для автоматического аннотирования // Программная инженерия. – 2014. – № 3. – С. 41-48.

[2] Алексеев А.А., Лукашевич Н.В. Комбинирование признаков для извлечения тематических цепочек в новостном кластере // Труды Института системного программирования РАН. – 2012. – Т. 23. – С. 257-276.

[3] Алексеев А.А., Лукашевич Н.В. Автоматическое извлечение сущностей на основе структуры новостного кластера // Искусственный интеллект и принятие решений. – 2011. – № 4. – С. 51-59.

Издания из списка SCOPUS:

[4] Alekseev A.A., Loukachevitch N.V. Use of Multiple Features for Extracting Topics from News Clusters // Proceedings of the Spring Researchers Colloquium on Databases and Information Systems. – 2012. – P. 3-11.

[5] Alekseev A.A., Loukachevitch N.V. The automatic retrieval of news entities based on the structure of a news cluster // Scientific and Technical Information Processing. – 2012. – Vol. 39, № 6. – P. 303-309.

[6] Alekseev A.A., Loukachevitch N.V. Automatic Entity Detection Based on News Cluster Structure // Proceedings of the International Workshop on Concept Discovery in Unstructured Data. – 2011. – P. 1-10.

Другие публикации:

[7] Алексеев А.А., Мальковский М.Г. Автоматическое аннотирование новостного кластера на основе тематического анализа // Тезисы докладов конференции «Тихоновские чтения». – М.: МГУ, 2013. – С. 55.

[8] Алексеев А.А. Тематическое представление новостного кластера как основа для автоматического аннотирования // Труды всероссийской конференции RCDL. – 2013. – С. 359-369.