Моделирование и анализ данных
2020. Том 10. № 3. С. 24–38
DOI: https://doi.org/10.17759/mda.2020100302
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2020. Vol 10. no. 3. pp. 24–38
DOI: https://doi.org/10.17759/mda.2020100302
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

УДК 004.912

Автоматическая кластеризация документов СМИ на основе анализа их смыслового содержания

Кан А.В.*
МАИ, г. Москва, Российская Федерация
ORCID: https://orcid.org/0000–0001–9410–406X
e-mail: kan_a@mail.ru

Козловская Я.Д.**
МАИ, г. Москва, Российская Федерация
ORCID: https://orcid.org/0000–0002–1780–5687
e-mail: yana_kozlovskaia@mail.ru

Кадушкин Н.А.***
МАИ, г. Москва, Российская Федерация
ORCID: https://orcid.org/0000–0002–0327–909X
e-mail: bbamrin@gmail.com

Хорошилов Ал-др.А.****
ГосНИИАС, г. Москва, Российская Федерация
ORCID: https://orcid.org/0000–0003–4885–3232
e-mail: aleksandr_khor@mail.ru

Для цитаты:
Кан А.В., Козловская Я.Д., Кадушкин Н.А., Хорошилов Ал-р А. Автоматическая кластеризация документов СМИ на основе анализа их смыслового содержания // Моделирование и анализ данных. 2020. Том 10. № 3. С. 24–38. DOI: https://doi.org/10.17759/mda.2020100302

***Кан Анна Владимировна,** доцент Московского авиационного института (национальный исследовательский университет), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000–0001–9410–406X, e-mail: kan_a@mail.ru

****Козловская Яна Дмитриевна,** магистрант Московского авиационного института (национальный исследовательский университет), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000–0002–1780–5687, e-mail: yana_kozlovskaia@mail.ru

*****Кадушкин Николай Алексеевич,** студент Московского авиационного института (национальный исследовательский университет), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000–0002–0327–909X, e-mail: bbamrin@gmail.com

******Хорошилов Александр Александрович,** инженер Государственного научно-исследовательского института авиационных систем (ГосНИИАС), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000–0003–4885–3232, e-mail: aleksandr_khor@mail.ru

В статье описывается решение проблемы автоматической кластеризации документов средств массовой информации (СМИ) на основе их смыслового анализа. Предлагаемое решение базируется на методах машинной грамматики, семантико-синтаксического и концептуального анализа текстов, а также на методах выявления понятийного состава коллекции документов и формализации смыслового содержания текстов. Разработанный алгоритм процесса кластеризации документов обеспечивает возможность его реализации в полностью автоматическом режиме без предварительного машинного обучения.

Ключевые слова: автоматическая кластеризация документов, машинная грамматика, семантико-синтаксический анализ текстов, концептуальный анализ текстов, актуальный концептуальный словарь.

ВВЕДЕНИЕ

Основная идея кластеризации коллекции документов заключается в разделении этой коллекции на группы (кластеры), совпадающих по смысловому содержанию. Это можно достигнуть на основе установления принципов сходства и различия документов. Сходство содержания документов можно установить наличием в них сходной по смыслу системы понятий, а различие определяется отсутствием таких понятий. При решении этой задачи необходимо из всей системы понятий, содержащейся в коллекции документов выделить ту систему значимых (ключевых) понятий, которая обладает достаточной смыслоразличительной способностью, обеспечивающей возможность разделения документов на кластеры с заданной полнотой и точностью. Выявление такой системы понятий можно обеспечить методами концептуального анализа лексического состава коллекции текстов. Такая система понятий должна составить основу семантического словаря, с помощью которого будет выявляться понятийная структура документов и строиться их поисковые образы.

Автоматическое создание словаря понятий по корпусу текстов достаточно подробно изложено в работе [8]. Отметим только то, что в основу концептуального анализа текста положены синтаксические, семантические и статистические методы анализа текстов.

Синтаксические методы позволяют выявить синтаксическую роль значимых слов и словосочетаний в предложении путем установления их синтаксической роли (принадлежность к словосочетаниям, являющимися в предложении группой подлежащего (субъекта), группой сказуемого (предиката) или группой дополнения (объекта)) [6–10].

Семантических методы позволяют выявить значимые в предметной области слова и словосочетания путем их соотнесения с элементами эталонных словарей или их формализованных семантических признаков с эталонными представлениями этих признаков [6–10].

Статистические методы позволяют путем назначения весовых коэффициентов установить состав значимых для коллекции документов понятий, выраженных словами и словосочетаниями, на основе анализа частот их появления в коллекции документов [2].

ТЕОРЕТИЧЕСКОЕ ОБОСНОВАНИЕ МЕТОДА КЛАСТЕРИЗАЦИИ

Предлагаемое решение задачи кластеризации базируется на концепции фразеологического концептуального анализа текстов [7; 8]. В соответствии с этой концепцией смысловое содержание текстов определяется системой текстовых понятий, проф. Г.Г. Белоногов определяет термин «“*понятие*” *как «социально значимый мыслительный образ, за которым в языке закреплено его наименование в виде отдельного слова или, значительно чаще, в виде устойчивого фразеологического словосочетания…»* [6], и их смысловых отношений, входящих в состав текстовых предложений. Смысл каждого предложения текста выражается через предикатно-актантную структуру.

Предикатно-актантная структура (ПАС) – описание предложений в виде понятий-актантов, выступающих в роли описываемых объектов, и понятий-предикатов, определяющих связи между объектами.

Предикат (*Predicate*), субъект (*Subject*) и объект (*Object*) представляются в виде шаблонов – обобщенных синтагм в нормализованном виде:

$$\mathrm{P} = \left\{ \mathrm{GS}_{norm_x} \mid x \in X \right\}, \tag{1}$$

$$S = \left\{ \mathrm{GS}_{norm_y} \mid y \in Y \right\}, \tag{2}$$

$$\mathrm{O} = \left\{ \mathrm{GS}_{norm_z} \mid z \in Z \right\}, \tag{3}$$

где x, y, z – нормированный набор индексов шаблона из словаря D обобщенных синтагм;

$X, Y, Z \in D$ – множества синтагм соответствующих элементов ПАС из словаря D;

GS_{norm} – обобщенная синтагма в нормализованном виде (лемма):

$$GS_{norm} = Lemma = \left\{ GS \mid n = 1, c = 1 \right\}, \tag{4}$$

где GS . – обобщенная синтагма, полученная при анализе всех допустимых грамматических признаков элементов ПАС предложения или словосочетания.

Смысл каждого предложения текста $SSen$ представляется через его предикатно-актантную структуру:

$$SSen = \sum_{i,j,k=1}^{I,J,K} \{S_i, P_j, O_k\} + \sum_{q=0}^{Q} R_q, \forall R \in \mathbb{R}, \tag{5}$$

где I, J, K, Q – счетные множества.

Идея кластеризации текстов СМИ по смысловому содержанию базируется на следующих постулатах и решениях:

1. Содержание каждого документа описывается системой значимых в предметной области понятий.
2. Основное смысловое содержание документа определяется системой его ключевых (значимых для предметной области) понятий.

3. Значимость понятий (веса смысловой значимости в предметной области) определяется статистическими, синтаксическими и семантическими критериями.
4. Понятия могут иметь различное текстовое представление, но при этом быть близким по своему смысловому содержанию.
5. Понятия в тексте связаны как парадигматическими (внеконтекстными), так и синтагматическими (контекстными) отношениями.
6. Синтагматические отношения разрешаются на основе семантико-синтаксического анализа текстов.
7. Парадигматические отношения разрешаются на основе использования таких семантических инструментов, как тезаурусы и онтологии.
8. Установление смыслового тождества понятий, представленных одним и тем же лексическим составом в различном синтаксическом контекстом окружении, разрешаются путем нормализации их лексического состава на уровне словоизменения или словообразования.

Общее решение задачи кластеризации обеспечивается следующими технологическими операциями.

1. Создание актуального словаря по текстам коллекции документов, полученного автоматически из частотного словаря путем исключения из него высокочастотной и низкочастотной частей этого словаря.
2. Повышения распознающей способности семантически значимой лексики в актуальном словаре, путем их сопоставления с наиболее значимыми понятиями предметной области.
3. Генерации на основе актуального словаря понятийных образов документов (ПОД), отражающих основное смысловое содержание документа.
4. Формирования кластеров путем последовательного сопоставления ПОДов документов-эталонов и ПОДов всех документов коллекции, и в случае превышения пороговых значений такие документы входят в состав кластера. После окончания каждого процесса формирования кластера этот документ-эталон и документы его кластера исключатся из «черного списка».
5. Процесс кластеризации завершается в случае полного опустошения «черного списка».
6. На завершающем этапе производится установление документа-центроида кластера (документа, смысловое содержание которого в наибольшей степени отражаются обобщенное содержание документов кластера).

РЕШЕНИЕ ЗАДАЧИ КЛАСТЕРИЗАЦИИ КОЛЛЕКЦИИ ДОКУМЕНТОВ

Рение задачи кластеризации можно условно разделить на несколько этапов, в рамках которых производятся технологические операции формализации представления содержанием отдельных документов, выявления смыслоразличающего понятийного состава коллекции документов иредставления обобщенного содержания документов кластеров в виде дайджеста кратких рефератов.

На первом этапе нужно определить множество слов и словосоч. етаний признакового пространства, по которым будет производиться установление смысловой схожести содержания документов коллекции. Для этого необходимо определить: по каким критериям нужно оценивать смысловую значимость каждого элемента признакового пространства и в соответствие с этими критериями назначить каждому слову или словосочетанию весовой коэффициент их смысловой значимости в признаковом пространстве.

Вачестве меры смысловой значимости слов и словосочетаний было решено использовать статистическую меру $TF-IDF$ (TF (term frequency) – частота слова).

TF – отношение числа вхождений наименования понятия общему числу наименований понятий документа. Таким образом, оценивается важность наименования понятия t_i в пределах отдельного документа.

$$TF\left(t,d\right)=\frac{n_t}{\sum_k n_k}, \tag{6}$$

где n_t . – число вхождений наименования понятия t в документ;

$\sum_k n_k$. – общее число наименований понятий в данном документе.

IDF – инверсия частоты, с которой некоторое наименование понятия встречается в документах коллекции. Для каждого уникального слова в пределах конкретной коллекции документов существует только одно значение *IDF*:

$$IDF(t,D)=\log\frac{\left|D\right|}{\left|\left\{d_i\in D\mid t\in d_i\right\}\right|}, \tag{7}$$

где $\left|D\right|$ – число документов в коллекции; $\left|\left\{d_i\in D\mid t\in d_i\right\}\right|$ – число документов из коллекции D, в которых встречается t (когда $n_t\neq 0$.).

Мера $TF-IDF$ является произведением двух сомножителей:

$$TF-IDF\left(t,d,D\right)=TF\left(t,d\right)\times IDF\left(t,D\right). \tag{8}$$

Большой вес в $TF-IDF$ получат наименования понятий с высокой частотой встречаемости в конкретном документе и с относительно низкой частотой в пределах всего корпуса текстов.

Но данная статистическая мера в явном виде не отражает смысловую составляющую наименований понятий. С этой целью была разработана система коррелирующих семантических весовых коэффициентов наименований понятий, восполняющих этот пробел. Обобщенная мера смысловой значимости наименований понятий M_7 с учетом этих коэффициентов вычисляются по формуле

$$M_7=(TF\cdot IDF)\cdot K_n\cdot K_z\cdot K_t\cdot K_w\cdot K_s\cdot K_f, \tag{9}$$

где:

K_n – коэффициент, учитывающий распознающую способность слов при их нормализации.

K_z – коэффициент, учитывающий вхождение в заголовки слов или словосочетаний.

K_t – коэффициент, учитывающий вхождение в тезаурус слов или словосочетаний.

K_w – коэффициент, учитывающий количество слов в словосочетании.

K_w – коэффициент, учитывающий синтаксическую роль слова или словосочетания в предложении.

K_f – коэффициент, учитывающий принадлежность понятия к фамильно-именной группе, бренду и др.

Для реализации процесса установления меры смысловой значимости наименований понятий коллекции документов было выполнено исследование по созданию актуального семантического словаря (АСК) для кластеризации текстов методами приближенного концептуального анализа текстов [6]. В рамках этого исследования был составлен частотный словарь наименований понятий и была сформирована его характеристическая таблица, устанавливающая закономерности покрытия текстов понятиями, содержащимися в коллекции документов, и рангами значений их частот. В табл. 1. представлена характеристическая таблица частотного словаря наименований понятий коллекции документов общим объемом 2013 документов.

Таблица 1

Характеристическая таблица частотного словаря наименований понятий коллекции документов

Ранг частот R_f	**Частота f**	**Кратность m**	**Накопленная частота A_f**	**Накопленная кратность A_m**	**Относительная накопл. част. C**
1	55255	1	55255	1	0,030403
2	42487	1	97742	2	0,053780
3	38668	1	136410	3	0,075056
4	38112	1	174522	4	0,096026
5	20854	1	195376	5	0,107500
6	18464	1	213840	6	0,117659
7	16508	1	230348	7	0,126742
8	16391	1	246739	8	0,135761
9	15827	1	262666	9	0,144470
10	14969	1	277535	10	0,152706
20	12311	1	413659	20	0,227604
30	7762	1	512051	30	0,281742
40	6866	1	584223	40	0,321452
60	5326	1	708632	61	0,389905
70	4183	1	756477	71	0,416230
80	3564	1	794862	81	0,437350
90	3217	1	828605	91	0,455917
100	2699	1	858548	101	0,472392
200	1214	1	1052304	209	0,579001
300	760	1	1160452	325	0,368506
400	518	1	1240985	456	0,682817
401	517	2	1242019	458	0,683386
500	366	2	1307201	609	0,719250
600	240	2	1377359	848	0,757853

Ранг частот R_f	Частота f	Кратность m	Накопленная частота A_f	Накопленная кратность A_m	Относительная накопл. част. C
700	137	27	1467679	1361	0,807549
794	43	31	1608826	3214	0,885211
806	31	47	1635882	3967	0,900098
827	10	500	1695851	7643	0,933094
828	9	426	1699685	8069	0,935204
829	8	692	1705221	8761	0,938250
830	7	681	1709988	9442	0,940873
831	6	1354	1718112	10796	0,945343
832	5	1971	1727967	12767	0,950765
833	4	3306	1741191	16073	0,958041
834	3	5053	1756350	21126	0,966382
835	2	13653	1783656	34779	0,981406
836	1	33793	1817449	68572	1,000000

Анализ содержания этой таблицы показывает, что число рангов частот наименований понятий коллекции документов содержит 836 рангов, а общее число наименований понятий в этой коллекции документов включает 1817449 понятий, из которых число разных понятий – всего 68572. При этом наименования понятий в коллекции распределены неравномерно. Так, например максимальная частота слова «проверка» имеет частоту 55255, в то же время более 32 тыс. слов и словосочетаний имеют частоту 1. В свете предлагаемых решений по созданию актуального семантического словаря было решено взять за основу словаря АСК данный частотный словарь, исключив из его состава высокочастотную и малочастотную части словаря. В процессе исследований было эмпирически установлено, что пороговым критерием для исключения высокочастотной лексики является значение частоты равное

$$\underline{f} = \frac{|\mathrm{D}|}{3} = \frac{2013}{3} = 671\,, \tag{10}$$

где $\underline{f}$ – высокочастотное пороговое значение; $|\mathrm{D}|$ – число документов в коллекции.

Этому значению частоты в наибольшей степени соответствует ранг частоты R_f со значением 300 (325 понятий, с покрытием понятий коллекции документов на 63,8 %).

При формировании низкочастотного порогового значения ранга частоты, определяющего число исключенных из состава частотного словаря малочастотных понятий, необходимо исходить из значения минимального объема кластера. Если принять, что минимальными объемом кластера может быть кластер с объемом не менее 10 документов, тогда нижнее пороговое значения частот принимает равной 10:

$$\overline{f} = |minclast| = 10\,, \tag{11}$$

где $\overline{f}$ – нижнее пороговое значение; $|minclast|$ – минимальное число документов в кластере. Этому критерию в данном случае будет соответствовать количество

разных понятий равное 7643 с покрытием понятий коллекции документов рангом частот более 10 на 93,3 %.

В этом случае объем АСК будет равен:

$$N = A_m\left[R_f(\overline{f})\right] - A_m\left[R_f(\underline{f})\right] = 7643 - 325 = 7318\,. \qquad (12)$$

Покрытие понятий коллекции документов понятиями АСК будет равен 30 % (93,3 %.-63,8 = 29,5 %). Такое значение покрытия АСК соответствует эмпирически установленному оптимальному значению покрытия автоматически создаваемых словарей АСК для различных коллекций текстов. Всем элементам словаря назначались весовые коэффициенты по формулам (8, 9).

Проведенные эксперименты показали недостаточную смыслоразличающую способность АСК, обусловленную тем фактом, что семантически значимые для заданной тематике однословные понятия и аббревиатуры имеют неприемлемо низкий весовой коэффициент. Поэтому было решено увеличить вес смысловой значимости тех однословным понятиям, которые входят в состав разработанного в процессе данного исследования тематического концептуального словаря (ТКС) и переназначить в словаре АСК их весовые коэффициенты. Те понятия, которые были идентифицированы с элементами словаря, включались в состав АСК и им присваивались новые весовые коэффициенты смысловой значимости в соответствии с формулой:

$$W_{POS} = f * \begin{cases} W_{POS} = N^2, \text{если } N > 1 \\ W_{POS} = 10, \text{если } N = \frac{1}{K_{Abr}}, \\ W_{POS} = 10, \text{если } N = K_{Sns} \end{cases} \qquad (13)$$

где f – частота; N – число слов в словосочетании; K_{Abr} – в словаре ТКС определено как аббревиатура; K_{Sns} – в словаре ТКС определено как значимое понятие предметной области.

Эксперименты, проведенные с модифицированным АСК с повышенными весовыми коэффициентами значимых понятий, показали его эффективность, увеличив вес ПОДов документов примерно на 15 %, и при этом существенно повысилась роль тех понятий, которые в наибольшей степени определяют основное смысловое содержание документов.

На втором этапе по текстам каждого документа коллекции был автоматически сформирован понятийный образ каждого документа (ПОД) коллекции. Формирование ПОДа осуществлялось путем выявления и идентификации в документе понятий, являющихся элементами АСК. При этом процесс идентификация понятий производилась на уровне словоизменительной нормализации слов, входящих в состав понятий.

Для каждого ПОДа документа вычислялась его семантическая характеристика (Под семантической характеристикой ПОДа понимается совокупность трех параметров: 1) число элементов Пода; 2) число высокозначимых понятий; 3) вес ПОДа., которая использовалась при формировании первоначального списка кластерообразую-

щих документов коллекции. В список кластерообразующих документов включались документы, семантическая характеристика которых превышала пороговое значение.

На третьем этапе выполнялся процесс автоматического формирования кластеров коллекции документов. Формирование каждого кластера производилось путем сопоставления элементов ПОДов кластерообразующего документа с ПОДами всей коллекции документов. В процессе такого сопоставления ПОДов вычислялось отношение суммы весов совпавших элементов ПОДа с весом ПОДа кластерообразующего документа. В зависимости от заданного порогового значения этого отношения принималось решение о включение документа коллекции в кластер данного кластерообразующего документа.

Для существенного уменьшения числа сравнений ПОДов документов коллекции в процессе кластеризации было принято следующее допущение:

Если процент совпадения весов ПОДов кластероообразующего документа и сравниваемого документа выше порогового (например, выше 30 %), то эти документы являются семантически близкими и в дальнейшем этот сравниваемый документ не должен использоваться в качестве кластероообразующего.

В рамках этих допущений было решено организовать два списка документов, изменяющихся в процессе кластеризации. Первый список («черный») должен включать документы, которые будут использовать в качестве кластерообразующих. Первоначально он сформирован на основе использования семантических характеристик ПОДа. В процессе кластеризации (формирования кластеров) из него должны исключаться те документы, которые уже использовались в качестве кластерообразующих и те документы, которые попали хотя бы в один из сформированных кластеров. Признаком завершения процесса кластеризации будет отсутствие документов в «черном» списке.

Одновременно для контроля процесса кластеризации будет создан постоянно пополняемый список («белый» список) документов, входящих хотя бы в один из сформированных кластеров. По завершению процесса кластеризации количество документов «белого» списка должно быть равно количеству документов коллекции.

Завершающим этапом процесса кластеризации являются операции нахождение центроида для каждого кластера, формирования названия кластера, а также проверка истинности процесса кластеризации. В процессе такой проверки должно быть установлено, что в каждом документе коллекции содержится обобщенное содержание кластера. Эту проверку можно осуществить путем формирования дайджеста кратких рефератов документов кластера. Все эти операции базируются на использовании автоматически создаваемого концептуального словаря кластера (КСК). Этот словарь строится для каждого кластера на основе анализа содержания документов, входящих в данный кластер. Поиск центроида осуществляется путем «взвешивания» вновь построенных ПОДов документов кластера по словарю КСК и удовлетворяющих двум условиям: 1) ПОД документ должен иметь максимальный вес, 2) источник документа должен быть включен в список верифицированных источников.

ТЕХНОЛОГИЧЕСКИЙ ПРОЦЕСС КЛАСТЕРИЗАЦИИ КОЛЛЕКЦИИ ДОКУМЕНТОВ

На основе предложенных решений был разработан обобщенный технологический процесс кластеризации, состоящий из следующих операций:

Этап 1. Конвертирование и формально-логический контроль каждого документа коллекции

1.1 Конвертирование исходных текстов во внутрисистемный формат.
1.2 Выполнение формально-логического контроля (ФЛК) текста.

Этап 2. Семантический анализ текста каждого документа коллекции

2.1 Графематический анализ текста.
2.2 Морфологический анализ слов текста.
2.3 Выделение наименований понятий текста.
2.4 Нормализация наименований понятий, выделенных из предложения.
2.5 Исключение незначимых слов и словосочетаний по словарю стоп-слов.
2.6 Построение списка формализованных наименований понятий документа с указанием частот их встречаемости.

Этап 3. Построение актуального концептуального словаря (АКС) для кластеризации коллекции документов

3.1 Создание массива всех формализованных наименований понятий с указанием частот их встречаемости в документах коллекции.
3.2 Построение частотного словаря формализованных наименований понятий коллекции документов.
3.3 Построение характеристической таблицы частотного словаря коллекции документов.
3.4 Определение верхнего и нижнего порогового значения частот словаря.
3.5 Формирование понятийного состава словаря АКС.
3.6 Назначение весового значения смысловой значимости наименованиям понятий словаря.
3.7 Корректировка значений весов понятий АКС по тематическому концептуальному словарю (ТКС).
3.8 Генерация машинного представления АКС.

Этап 4. Построение понятийного образа документа (ПОДа) коллекции

4.1 Выявление в тексте наименований понятий по словарю АКС.
4.2 Формирования ПОДа документа.
4.3 Назначение каждому элементу ПОДа весовых характеристик.

Этап 5. Кластеризация коллекции текстов

5.1 Получение информации о семантических характеристиках ПОДов документов коллекции.
5.2 Формирование первоначального «черного» списка документов коллекции.
5.3 Последовательное сопоставление ПОДов «черного» списка документов с документами коллекции.

5.4 Ведение «черного» и «белого» списков документов коллекции в процессе кластеризации.
5.5 Формирование кластеров на основе анализа результатов сопоставление ПОДов документов.

Этап 6. Установление центроида кластера

6.1 Выделение наименований понятий из документов кластера.
6.2 Нормализация наименований понятий, выделенных из предложения.
6.3 Исключение незначимых слов и словосочетаний по словарю стоп-слов.
6.4 Построение списка формализованных наименований понятий документа с указанием частот их встречаемости.
6.5 Формирование понятийного состава концептуального словаря кластера (КСК).
6.6 Назначение значений коэффициентов смысловой значимости наименованиям понятий словаря КСК.
6.7 Корректировка значений весов понятий словаря КСК по ТКС.
6.8 Генерация машинного представления КСК.
6.9 Формирования ПОДа документов кластера по словарю КСК.
6.10 Определение веса ПОДов документа кластера.
6.11 Определение центроида кластера.

Этап 7. Генерация обобщенного содержания кластера в виде дайджеста рефератов

7.1 Определение местоположения наименований понятий в предложениях документов кластера.
7.2 Вычисление суммарных весов предложений документа.
7.3 Установление заданных N предложений с максимальным весом.
7.4 Построение краткого реферата каждого документа, состоящего из заголовка и N заданных предложений с максимальным весом.
7.5 Формирование дайджеста обобщенных рефератов документов кластера.

На основе предложенных теоретических решений было разработано экспериментальное программное обеспечение. Разработанная программно-технологическая схема процесса кластеризации документов приведена на рис. 1. Представленная схема состоит из программного модуля хранения декларативных средств и семи последовательно выполняемых программных модулей:

1. Модуль конвертирования и ФЛК коллекции текстов.
2. Модуль семантического анализ текста документа.
3. Модуль построения АКС.
4. Модуль построения ПОДа.
5. Модуль кластеризации коллекции текстов.
6. Модуль установления центроида кластера.
7. Модуль генерации обобщенного содержания кластера.

В свою очередь, как видно на рис. 1, каждый программный модуль состоит из нескольких функциональных модулей, выполняющих конкретную целевую функцию. Реализация файлового обмена между программными модулями и модулями позволяет в полной мере контролировать процесс кластеризации на различных его этапах.

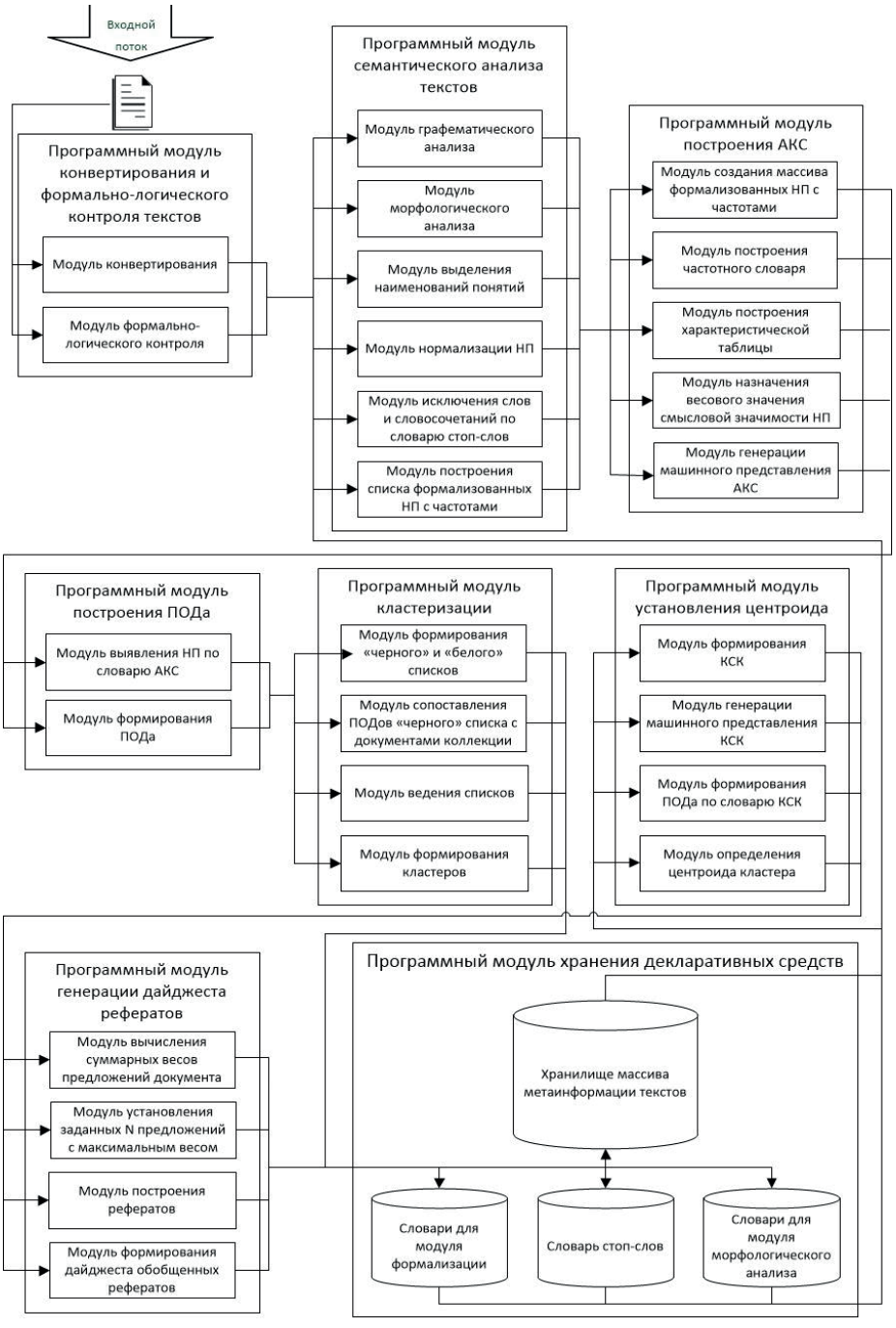

Рис. 1. Система кластеризации документов

ЗАКЛЮЧЕНИЕ

В статье описано решение задачи автоматической кластеризации документов СМИ на основе их смыслового анализа. Предлагаемое решение базируется на методах машинной грамматики, семантико-синтаксического и концептуального анализа текстов, а также на методах выявления понятийного состава коллекции документов и формализации смыслового содержания текстов. Автоматическое выявление в текстах наименований понятий (сущностей) базируется на методах концептуального анализа текстов и методике вычисления меры их смысловой значимости, использующей статистические, синтаксические и семантические признаки. На основе предложенных методов разработаны алгоритмы и программное обеспечение автоматической кластеризации документов СМИ на основе их смыслового анализа. Проведенные эксперименты показали работоспособность предложенных решений автоматической кластеризации коллекции документов и возможность их использования в промышленных реализациях.

Литература

1. *Богатырев М.Ю.* Извлечение фактов из текстов естественного языка с применением концептуальных графовых моделей // Известия ТулГУ. Технические науки. – 2016. – № 7. – Ч. 1.
2. Современные технологии обработки естественного языка в задачах стратегического управления / Виноградов А.Н. [и др.] // Технологическая перспектива в рамках евразийского пространства: новые рынки и точки экономического роста. / Власова Н.А., Куршев Е.П., Подобряев А.В. – СПб.:Центр научно-информационных технологий «Астерион», – 2018.
3. *Ермаков А.Е.* Автоматическое извлечение фактов из текстов досье: опыт установления анафорических связей [Электронный ресурс] // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции «Диалог’2007». – М. : Наука, – 2007.
4. Автоматическое создание формализованного представления смыслового содержания неструктурированных текстовых сообщений СМИ и социальных сетей / Хорошилов Ал-др. А. [и др.] // Системы высокой доступности / Никитин Ю.В., Хорошилов Ал-ей. А., Будско В.И., – 2014. – Т. 10., № 3.
5. *Helbig H.* Knowledge representation and the semantics of natural language. – Berlin: Springer, – 2006.
6. *Белоногов Г.Г., Гиляревский Р.С., Хорошилов А.А.* Проблемы автоматической смысловой обработки текстовой информации // Научно-техническая информация. Сер. 2. Информационные процессы и системы / Всероссийский институт научной и технической информации РАН. – 2012, № 11. – С. 24–28.
7. Средства машинной грамматики русского языка (по Г.Г. Белоногову) / Аблов И.В. [и др.] // Научно-техническая информация / Козичев В.Н., Ширманов А.В., Хорошилов Ал-др А., Хорошилов Ал-ей А., Сер. 2, – 2018. № 6.
8. *Калинин Ю.П., Хорошилов Ал-др. А., Хорошилов Ал-ей. А.* Современные технологии автоматизированной обработки текстовой информации // Системы высокой доступности, – 2015. – Т. 11, № 2.

Automatic Clustering of Mass Media Documents Based on the Analysis of Their Semantic Content

Anna V. Kan*
Moscow Aviation Institute (National Research University), Moscow, Russia
ORCID: https://orcid.org/0000–0001–9410–406X
e-mail: avkan@nrczh.ru

Yana D. Kozlovskaya**
Moscow Aviation Institute (National Research University), Moscow, Russia
ORCID: https://orcid.org/0000–0002–1780–5687
e-mail: yana_kozlovskaia@mail.ru

Nikolay A. Kadushkin****
Moscow Aviation Institute (National Research University), Moscow, Russia
ORCID: https://orcid.org/0000–0002–0327–909X
e-mail: bbamrin@gmail.com

Aleksandr A. Khoroshilov*****
State Scientific Research Institute of aircraft systems (GosNIIAS), Moscow, Russia
ORCID: https://orcid.org/0000–0003–4885–3232
e-mail: aleksandr_khor@mail.ru

The article describes the solution to the problem of automatic clustering of media documents based on the analysis of their semantic analysis. The proposed solution is based on the methods of machine grammar, semantic-syntactic and conceptual analysis of texts, as well as methods for identifying the conceptual composition of a collection of documents and formalizing the semantic content of texts. The developed algorithm of the document clustering process provides for the possibility of its implementation in a fully automatic mode without prior machine learning.

Keywords: automatic clustering of documents, machine grammar, semantic-syntactic analysis of texts, conceptual analysis of texts, actual conceptual vocabulary.

For citation:
Kan A.V., Kozlovskaya Y.D., Kadushkin N.A., Khoroshilov Al-dr.A. Automatic Clustering of Mass Media Documents Based on the Analysis of Their Semantic Content. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*, 2020. Vol. 10, no. 3, pp. 24–38. DOI: https://doi.org/10.17759/mda.2020100302 (In Russ., abstr. in Engl.).

***Anna V. Kan,** associate professor of the Institute of Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: https://orcid.org/0000–0001–9410–406X, e-mail: avkan@nrczh.ru

****Yana D. Kozlovskaya,** student of the Institute of Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: https://orcid.org/0000–0002–1780–5687, e-mail: yana_kozlovskaia@mail.ru

*****Nikolay A. Kadushkin,** student of the Institute of Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: https://orcid.org/0000–0002–0327–909X, e-mail: bbamrin@gmail.com

******Aleksandr A. Khoroshilov,** engineer of State Scientific Research Institute of aircraft systems (GosNIIAS), Moscow, Russia, ORCID: https://orcid.org/0000–0003–4885–3232, e-mail: aleksandr_khor@mail.ru

References

1. Bogatyrev M. Yu. Izvlechenie faktov iz tekstov estestvennogo yazyka s primeneniem kontseptual'nykh grafovykh modelei [Fact extraction from natural language texts with conceptual graph models]. Izvestiya TulGU. Tekhnicheskie nauki. – 2016. – № 7. – Ch. 1.
2. Vinogradov A.N. [i dr.] Sovremennye tekhnologii obrabotki estestvennogo yazyka v zadachakh strategicheskogo upravleniya [Modern technologies of natural language processing in strategic management tasks]. Tekhnologicheskaya perspektiva v ramkakh evraziiskogo prostranstva: novye rynki i tochki ekonomicheskogo rosta. – SPb.:Tsentr nauchno-informatsionnykh tekhnologii "Asterion", 2018.
3. Ermakov A.E. Avtomaticheskoe izvlechenie faktov iz tekstov dos'e: opyt ustanovleniya anaforicheskikh svyazei [Elektronnyi resurs] [Automatic extraction of facts from dossier texts: an experience of establishing anaphoric connections]. Komp'yuternaya lingvistika i intellektual'nye tekhnologii: trudy Mezhdunarodnoi konferentsii «Dialog'2007». – Moscow. : Nauka, 2007.
4. Khoroshilov Al-dr. A. [i dr.] Avtomaticheskoe sozdanie formalizovannogo predstavleniya smyslovogo soderzhaniya nestrukturirovannykh tekstovykh soobshchenii SMI i sotsial'nykh setei [Automatic creation of a formalized representation of the semantic content of unstructured text messages of the media and social networks]. Sistemy vysokoi dostupnosti, № 3, Vol. 10, 2014.
5. Helbig H. Knowledge representation and the semantics of natural language. – Berlin: Springer, 2006.
6. Belonogov G.G., Gilyarevskii R.S., Khoroshilov A.A. Problemy avtomaticheskoi smyslovoi obrabotki tekstovoi informatsii [Problems of automatic semantic processing of text information]. Nauchno-tekhnicheskaya informatsiya. Ser. 2. Informatsionnye protsessy i sistemy / Vserossiiskii institut nauchnoi i tekhnicheskoi informatsii RAN. – 2012, № 11. – pp. 24–28.
7. Ablov I.V. [i dr.] Sredstva mashinnoi grammatiki russkogo yazyka (po G.G. Belonogovu) [Means of machine grammar of the Russian language (according to G.G. Belonogov)]. Nauchno-tekhnicheskaya informatsiya. Ser. 2, № 6, 2018.
8. Kalinin Yu.P., Khoroshilov Al-dr. A., Khoroshilov Al-ei. A. Sovremennye tekhnologii avtomatizirovannoi obrabotki tekstovoi informatsii [Modern technologies for automated processing of text information]. Sistemy vysokoi dostupnosti, № 2, Vol. 11, 2015.