УДК 004.02:004.032.26

Алыгулиев Р.М.
Институт Информационных Технологий НАНА, Баку, Азербайджан
aramiz@iit.ab.az a.ramiz@science.az

РЕФЕРИРОВАНИЕ НАБОРА ДОКУМЕНТОВ ЧЕРЕЗ КЛАСТЕРИЗАЦИЮ И РАНЖИРОВАНИЕ ПРЕДЛОЖЕНИЙ

Предложен подход составления обзорных рефератов по набору документов, с выявлением тематических разделов и извлечением информативных предложений. Тематические разделы определены с помощью кластеризации предложений, а информативные предложения – применением алгоритма ранжирования. Показано, что результаты реферирования зависят от метода кластеризации, алгоритма ранжирования и меры подобий. Эксперименты на открытых корпусах DUC2001 и DUC2002 доказали, что предложенные методы кластеризации и алгоритм ранжирования показывают лучшие результаты, чем известный метод k-средних и алгоритмы ранжирования PageRank и HITS.

Ключевые слова: *многодокументное реферирование, кластеризация и ранжирование предложений.*

Введение

В зависимости от количества реферируемых документов-источников рефераты могут быть однодокументными (или монографическими) и многодокументными (или обзорными). Рефераты, составленные по одному документу, называются однодокументными, а рефераты, составленные по нескольким документам на одну тему, являются многодокументными. Однодокументное реферирование может сжать только один документ, представив его в более краткой форме, тогда как многодокументное реферирование может сжать набор документов в один реферат. Многодокументное реферирование можно рассматривать как расширение однодокументного реферирования. Его используют для сжатого описания информации, содержащейся в группе документов, что облегчает пользователям понимание набора документов, так как оно объединяет и интегрирует информацию, содержащуюся в документах, осуществляет синтез и обнаружение знаний, которое может быть использовано для приобретения знаний [1–5].

Сущность многодокументного реферирования может быть формально изложена таким образом. Дан набор документов, относящихся к одной и той же теме или событию. Нужно извлечь содержание из набора документов, относящихся к теме или событию, удаляя при этом излишнюю информацию и принимая во внимание сходные и различающиеся моменты в содержании, и представить пользователю в сжатой форме самую важную информацию таким образом, чтобы она отвечала нуждам пользователя или приложения [1, 2, 3].

Автоматическое реферирование набора документов привлекает внимание как решение проблемы информационной перегрузки и помогает пользователям в сканировании большого количества документов, представляющих для них интерес. Это важная функция, которая должна быть в наличии в больших цифровых библиотечных системах, системах поиска информации и поисковых машинах web,

где главной проблемой для пользователей является нахождение слишком большого числа документов [2, 4–6].

Существуют два подхода к автоматическому реферированию: экстрагирование и абстрагирование [2, 5]. Первый подход ориентирован на извлечение фрагментов (обычно предложений, отсюда общее обозначение подхода – sentence extraction) исходного документа, из которых и составляется реферат. Второй подход предполагает использование более изощренных методов лингвистического и семантического анализов. В данном случае обычно говорят о генерации реферата (summary generation) на основе семантического представления текста. Абстракция включает в себя генерацию новых предложений на основе информации, извлеченной из источника (источников). При абстрагировании требуются информационное слияние, сжатие и переформулировка предложений. Результат методов, основанных на извлечении предложений, далек от идеала – связного реферата, составленного квалифицированным специалистом, однако он лишен субъективизма. Главным ограничением обоих подходов является требование сжатия. Объем реферата, в зависимости от пользования, должен составлять от 5 до 30% исходного текста [2,4, 5].

Извлечение предложений – один из широко используемых подходов в многодокументном реферировании, в котором из документа (или документов) извлекаются наиболее значимые предложения и представляются как реферат. Данная работа тоже концентрируется на экстрактивное многодокументное реферирование, которое состоит из двух этапов: кластеризации и ранжирования. На первом этапе осуществляется кластеризация предложений для выявления тематических разделов, на втором этапе производится ранжирование предложений с целью извлечения наиболее информативных предложений.

Реферирование набора документов через кластеризацию и ранжирование предложений

В последние годы в многодокументном реферировании широко применяются алгоритмы ранжирования, основанные на теории графов [7–13]. Они порождены общеизвестными алгоритмами PageRank и HITS (Hyperlink Induced Topic Search) [14] или алгоритмом manifold-ranking [15]. Алгоритмы ранжирования генерируют связи между предложениями и на их основе анализируют структуру документа. Затем знания, полученные из структуры документов, применяются к извлечению предложений. Для вычисления относительной значимости предложений Erkan и др. [8] предложили стохастический метод, где важность предложений измеряется центральностью собственных векторов в графовом представлении предложений. Эта работа на основе графа подобия вводит три разных метода вычисления центральности предложений: степень центральности, LexRank и непрерывный LexRank. В [9] предложена тематико-чувствительная (topic-sensitive) или настраиваемая (biased) версия метода LexRank. Основное преимущество этого метода в том, что он обладает единственным параметром-настройкой, с помощью которой можно эффективно определить, насколько результирующий реферат будет общим или запросоориентированным. В отличие от обучаемых методов, он не требует большого количества обучающих данных – рефератов. В алгоритме GSPSummary [10] ранжирование предложений производится на основе их

значимости, полученной с помощью «персонализированного» LexRank. В этом алгоритме минимальная избыточность в реферате достигается путем разбиения документа на подтемы (кластеры), где в качестве глобального признака берется «персонализированный» LexRank-вектор. Для создания запросоориентированного многодокументного реферата высокого качества алгоритм PPRSum [11] в одной унифицированной структуре объединяет несколько видов информации. Сначала, используя наивную Байесовскую модель, он обучает модель значимости глобальных признаков предложения. Затем, используя запрос, для каждого корпуса создает модель релевантности. В конце, с использованием обеих моделей – модели значимости и модели релевантности, для каждого предложения в корпусе вычисляется персонализированная априорная вероятность. С помощью персонализированной априорной вероятности, в зависимости от отношений между предложениями в корпусе, производится ранжирование предложений. Наконец, во избежание избыточности в реферате, на каждое предложение налагается штраф за избыточную информацию. Реферат создается выбором предложений, где учитываются их высокая информационная новизна и близость к запросу. Предполагая, что реферат и представление темы взаимно дополняющие, Zhang и др. [12] предложили адаптивную модель реферирования AdaSum. Новый метод ранжирования, iSpreadRank [13], множество документов моделирует как сети подобия предложений. Основанный на такой сетевой модели алгоритм iSpreadRank, используя теорию распространяющейся активации, формулирует общую концепцию, взятую из социальных сетей. В этом методе значимость узла (т.е. предложения) в сети определяется не только числом узлов, с которыми он соединен, но и важностью соединенных с ним узлов. Алгоритм рекурсивно перевзвешивает важность предложений, с распространением их веса на всю сеть, для оценки значимости других предложений.

Алгоритм manifold ranking [15] описывает связную структуру документа при помощи матриц. Алгоритм учитывает как зависимости между предложениями внутри одного документа, так и зависимости между предложениями всего набора документов. Составление обзорных рефератов с помощью алгоритма manifold ranking состоит из двух этапов:

1) вычисление ранга каждого предложения. Этим решается задача ранжирования всех предложений в соответствии с их «близостью» заданной теме;

2) применение алгоритма отсечения предложений, наиболее похожих на те, что уже включены в обзорный реферат. Этим решается исключение из обзорного реферата одинаковых или близких предложений, т.е. проблема избыточности в обзорном реферате.

В результате с учетом коэффициента сжатия некоторое количество предложений с наибольшим рангом выбирается для включения в реферат. В работе [16] проведен анализ возможности использования алгоритма manifold ranking для русского языка.

Для ранжирования предложений авторы статьи [17] предложили две модели. Первая модель, ClusterCMRW (Cluster-based Conditional Markov Random Walk Model), на основе информации, полученной через кластеризацию предложений, строит граф связей. Вторая модель, ClusterHITS (Cluster-based HITS Model),

кластеры и предложения рассматривает как «концентраторы» и «авторитеты» алгоритма HITS.

Если документы затрагивают несколько тем, то проблема создания реферата, охватывающего все темы, становится очень трудной. Трудность заключается в том, что алгоритмы ранжирования [14] не могут выявить тематические разделы. Для решения этой проблемы в работах [18, 19] предложены подходы, позволяющие выявить скрытые тематические разделы. Идея метода, предложенного в данной работе, состоит в том, чтобы сначала найти тематические разделы набора документов, т.е. группы предложений, относящихся к одной подтеме. После кластеризации, с учетом степени значимости кластеров, применяя алгоритм ранжирования, извлечь информативные предложения.

Кластеризация предложений – первый этап. Задан набор документов $\mathbf{D} = [D_1, D_2, ..., D_d]$. Каждый документ D_i представляется как множество предложений $D_i = [S_{i,1}, S_{i,2}, ..., S_{i,n_i}]$, где S_{ij} обозначает j-е предложение в D_i, n_i – число предложений в документе D_i. Пусть $T = [t_1, t_2, ..., t_m]$ представляет все слова, встречающиеся в коллекции $\mathbf{D}$, где m – общее число слов. Пусть $\mathbf{S} = [S_1, S_2, ..., S_n]$ совокупность предложений в коллекции $\mathbf{D}$, которую требуется разбить на непересекающиеся кластеры $\mathbf{C} = [C_1, ..., C_k]$. Кластеризация предложений осуществляется по методу [19]:

$$\mathcal{F} = \frac{2}{\frac{1}{\mathcal{P}} + \frac{1}{\mathcal{R}}} = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P} + \mathcal{R}} \to \max . \tag{1}$$

Функции $\mathcal{P}$ и $\mathcal{R}$ [18,19] соответственно обеспечивают компактность

$$\mathcal{P} = \left(\sum_{p=1}^{k} \left| C_p \right| \sum_{S_i, S_j \in C_p} sim(S_i, S_j) \right)^{-1} \to \max \tag{2}$$

и отделимость кластеров

$$\mathcal{R} = \sum_{p=1}^{k-1} \sum_{q=p+1}^{k} \sum_{S_i \in C_p} \sum_{S_j \in C_q} \left| C_p \right\| C_q \left| sim(S_i, S_j) \right. \to \max , \tag{3}$$

где мера подобия $sim(S_i, S_j)$ равна расширенной мере Jaccard [20]

$$sim(S_i, S_l) = sim_{\text{Jaccard}}(S_i, S_l) = \frac{\left| S_i \bigcap S_l \right|}{\left| S_i \right| + \left| S_l \right| - \left| S_i \bigcap S_l \right|} \tag{4}$$

или мере, основанной на попарной близости слов [18,19]:

$$sim(S_i, S_j) = sim_{\text{NGD}}^{\text{weight}}(S_i, S_j) = w \cdot sim_{\text{NGD}}^{\text{global}}(S_i, S_j) + (1 - w) \cdot sim_{\text{NGD}}^{\text{local}}(S_i, S_j) . \tag{5}$$

В формуле (5) «локальная» и «глобальная» меры подобия определяются так [18, 19]:

$$sim_{\text{NGD}}^{\text{local}}(S_i, S_j) = \frac{\sum_{t_k \in S_i} \sum_{t_l \in S_j} \text{NGD}^{\text{local}}(t_k, t_l)}{m_i m_j} , \tag{6}$$

$$sim_{\text{NGD}}^{\text{global}}(S_i, S_j) = \frac{\sum_{t_k \in S_i} \sum_{t_l \in S_j} \text{NGD}^{\text{global}}(t_k, t_l)}{m_i m_j} , \tag{7}$$

$$\mathrm{NGD}^{\mathrm{global}}(t_k,t_l)=\frac{\max\{\log(f_k^{\mathrm{global}}),\log(f_l^{\mathrm{global}})\}-\log(f_{kl}^{\mathrm{global}})}{\log d-\min\{\log(f_k^{\mathrm{global}}),\log(f_l^{\mathrm{global}})\}}, \quad (8)$$

$$\mathrm{NGD}^{\mathrm{local}}(t_k,t_l)=\frac{\max\{\log(f_k^{\mathrm{local}}),\log(f_l^{\mathrm{local}})\}-\log(f_{kl}^{\mathrm{local}})}{\log n-\min\{\log(f_k^{\mathrm{local}}),\log(f_l^{\mathrm{local}})\}}, \quad (9)$$

где f_k^{global} – число документов, в которых встречается слово t_k, f_{kl}^{global} – число документов, в которых встречаются оба слова t_k и t_l, а d – число документов в наборе $\mathbf{D}$, $n=\sum_{i=1}^{n} n_i$ – общее число предложений в наборе $\mathbf{D}$; f_k^{local} – число предложений, в которых встречается слово t_k, а f_{kl}^{local} – число предложений, в которых встречаются оба слова t_k и t_l, $w\in[0,1]$ – параметр взвешивания, определяющий относительные вклады «локальной» (6) и «глобальной» (7) мер подобия в общую меру подобия (5).

Ранжирование предложений – второй этап. Традиционные алгоритмы ранжирования не отличают межкластерные ребра от внутрикластерных и не учитывают степень значимости кластера. Поэтому предлагается такая модификация алгоритма WICER [21], которая межкластерные ребра отличает от внутрикластерных и принимает во внимание степень важности кластеров. Основная идея алгоритма WICER заключается в следующем:

1) межкластерное ребро имеет более высокую значимость, чем внутрикластерное;

2) каждый кластер взвешен на основе его степени важности, т.е. степени охвата контента всей коллекции документов.

С учетом последних высказываний алгоритм WICER [21] модифицируем в таком виде:

$$PR_p(S_j)=\frac{(1-d)}{n}+d\left(1+\frac{n_p}{k}\right)\sum_{q=1}^{k}W_q\left(\sum_{S_i\in C_q^i}\frac{sim(S_i,S_j)}{\sum_{S\in \mathrm{adj}[S_j]}sim(S,S_j)}w_{pq}PR_q(S_i)\right), \quad (10)$$

где

- n_p – число кластеров, связанных с кластером C_p;
- $PR_p(S_i)$ – ранг предложения S_i кластера C_p;
- C_q^i – множество предложений в кластере C_q, которое связано с предложением S_i;
- W_q – вес кластера C_q, который равен мере подобия между кластером и набором документов, $W_q=sim(C_q,\mathbf{D})$;
- w_{pq} – вес ребра из кластера C_p в кластер C_q, $w_{pq}=\begin{cases}\alpha, \text{если } p=q\\ \beta, \text{если } p\neq q\end{cases}$, где $\beta>\alpha$, т.е. полагается, что межкластерное ребро более значимое, чем внутрикластерное (в экспериментах их значения установили так: $\alpha=1$ и $\beta=1{,}3$);
- $\mathrm{adj}[S_i]$ – множество предложений, связанных с предложением S_i.

Как можно было заметить, алгоритм (10) при ранжировании предложений учитывает и количество связанных с ним кластеров (множитель $1+n_p/k$).

Эксперименты и анализ

Чтобы показать эффективность предложенного подхода, будем сравнивать его с подобным методом [17], где авторы, применяя известные методы кластеризации (*k*-средних, агломеративный и дивизимный), кластеризуют предложения, потом с помощью алгоритмов PageRank и HITS [14] ранжируют их. Метод, использующий алгоритм PageRank, назван ClusterCMRW, а использующий алгоритм HITS – ClusterHITS.

Для измерения степени согласия между рефератами, составленными разными системами, наиболее распространенной мерой является ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [22]. Со стороны DUC [23] она принята как официальная мера для оценки систем реферирования текстов. Она измеряет, насколько автоматически созданный реферат перекрывается с эталонными рефератами, используя при этом статистики совпадения N-грамм. В зависимости от стратегии оценки мера ROUGE классифицируется в следующие: ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S и ROUGE-SU. ROUGE-N представляет собой обобщенную статическую меру, выражающую, какой процент N-грамм, входящих в эталонный реферат, попадает в автоматически созданный реферат [22]:

$$\text{ROUGE}-\text{N}=\frac{\sum_{S\in \text{Summ}_{\text{ref}}}\sum_{\text{N}-\text{gram}\in S} Count_{\text{match}}(\text{N-gram})}{\sum_{S\in \text{Summ}_{\text{ref}}}\sum_{\text{N-gram}\in S} Count(\text{N-gram})}, \qquad (11)$$

где N обозначает длину N-грамм, $Count_{\text{match}}(\text{N-gram})$ – максимальное число совпадающих N-грамм в автоматически созданном реферате и в эталонных рефератах, $Count(\text{N-gram})$ – число N-грамм в эталонных рефератах. ROUGE-L – самая длинная общая подпоследовательность (longest common subsequence – LCS) слов в автоматически созданном и эталонном рефератах. ROUGE-W при оценке использует взвешенную LCS (вес равен 1,2). ROUGE-S использует отношение перекрывания skip-bigram между автоматически созданным и эталонным рефератами. ROUGE-SU является расширением ROUGE-S с добавлением униграмм.

Чтобы быть уверенным в адекватности предложенного подхода, в экспериментах были выбраны одни и те же наборы данных – DUC2001 и DUC2002, метод кластеризации – *k*-средних, мера подобия – мера косинуса и меры оценки – ROUGE-1, ROUGE-2 и ROUGE-W. Для каждого кластера был создан реферат длиной в 100 слов. В наших экспериментах в том числе были изучены влияния меры подобия и метода кластеризации на результат реферирования. С этой целью были сравнены мера косинуса, расширенная мера Jaccard (4) и мера, основанная на попарной близости слов (5).

Влияние метода кластеризации на точность нашего алгоритма было изучено сравнением метода $\mathcal{F}$ (1) с методом *k*-средних. Результаты эксперимента на наборах данных DUC2001 и DUC2002 показаны соответственно в таблицах 1 и 2. В таблицах наш алгоритм обозначен через ClusterWICER. Каждая таблица разбита на две части: верхнюю и нижнюю части. Верхняя часть сравнивает результаты разных

алгоритмов, а нижняя часть – результаты алгоритма ClusterWICER, полученные при разных мерах подобия, где в качестве метода кластеризации выбран $\mathcal{F}$ (1).

Таблица 1

ROUGE-показатели методов на наборе данных DUC2001

Алгоритм реферирования	Мера подобия	Метод кластеризации	ROUGE-1	ROUGE-2	ROUGE-W
ClusterWICER	$sim_{\cos}$	*k*-средних	0,37981	0,07173	0,11439
ClusterCMRW	$sim_{\cos}$	*k*-средних	0,35824	0,06458	0,10770
ClusterHITS	$sim_{\cos}$	*k*-средних	0,35756	0,05944	0,10771
ClusterWICER	sim_{Jaccard}	$\mathcal{F}$	0,38327	0,07749	0,11623
ClusterWICER	$sim_{\text{NGD}}^{\text{weight}}$	$\mathcal{F}$	0,38840	0,07986	0,11978
ClusterWICER	$sim_{\cos}$	$\mathcal{F}$	0,38136	0,07514	0,11596

Таблица 2

ROUGE-показатели методов на наборе данных DUC2002

Алгоритм реферирования	Мера подобия	Метод кластеризации	ROUGE-1	ROUGE-2	ROUGE-W
ClusterWICER	$sim_{\cos}$	*k*-средних	0,38925	0,08859	0,12776
ClusterCMRW	$sim_{\cos}$	*k*-средних	0,38221	0,08321	0,12362
ClusterHITS	$sim_{\cos}$	*k*-средних	0,37643	0,08135	0,12141
ClusterWICER	sim_{Jaccard}	$\mathcal{F}$	0,39423	0,09159	0,13253
ClusterWICER	$sim_{\text{NGD}}^{\text{weight}}$	$\mathcal{F}$	0,39734	0,09286	0,13609
ClusterWICER	$sim_{\cos}$	$\mathcal{F}$	0,39281	0,09027	0,12984

Из анализа таблиц получим следующие основные результаты:

- алгоритм ClusterWICER по всем ROUGE-показателям превышает алгоритмы ClusterCMRW и ClusterCMRW (рис. 1 и 2);
- ROUGE-показатели алгоритмов на наборе DUC2002 лучше, чем на наборе DUC2001;
- ClusterCMRW показывает лучший результат, чем алгоритм ClusterHITS;

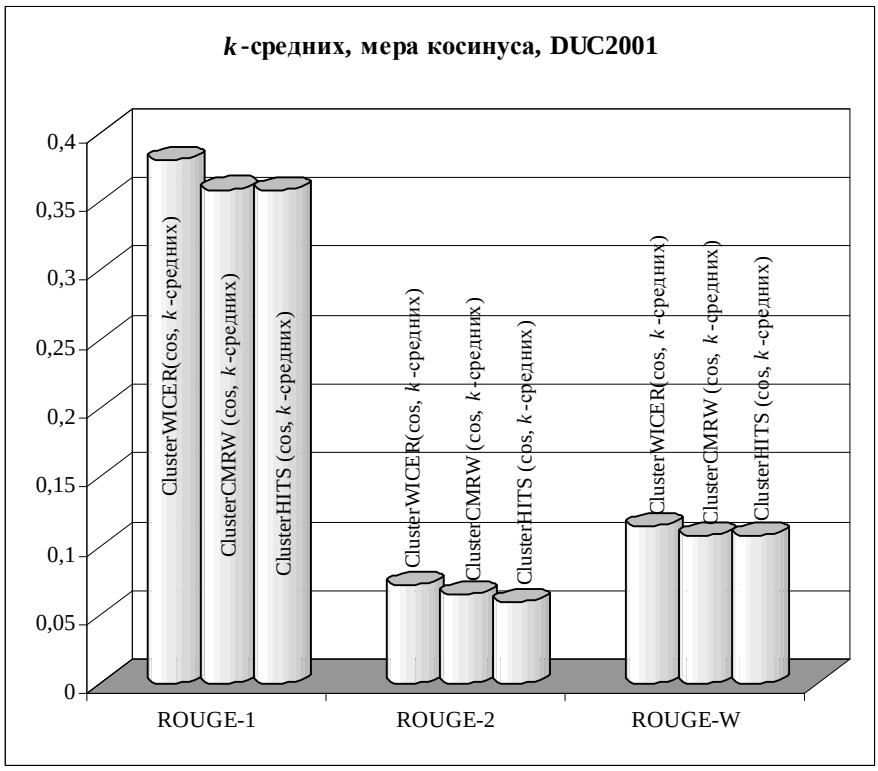

Рис.1. Сравнение алгоритмов реферирования на наборе DUC2001 (*k*-средних и мера косинуса).

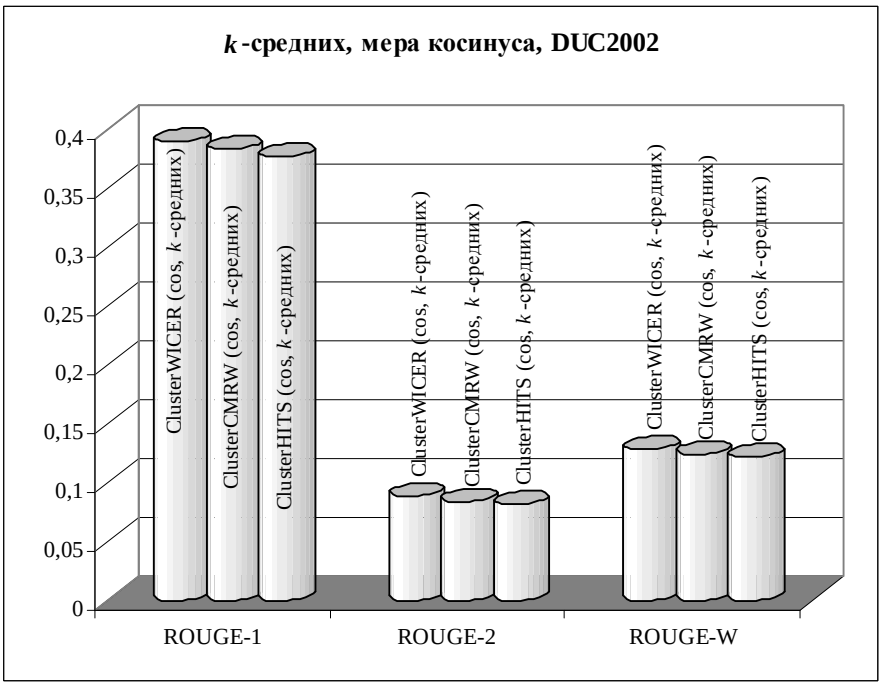

Рис.2. Сравнение алгоритмов реферирования на наборе DUC2002 (*k*-средних и мера косинуса).

- сравнение ClusterWICER($sim_{\cos}$,*k*-средних) с ClusterWICER($sim_{\cos}$,$\mathcal{F}$) демонстрирует, что последний имеет лучший результат, где в скобках показаны мера подобия и метод кластеризации (рис. 3 и 4). Другими словами, функция $\mathcal{F}$ обладает преимуществом над методом *k*-средних;

- среди алгоритмов ClusterWICER($sim_{\text{Jaccard}}, \mathcal{F}$), ClusterWICER($sim_{\text{NGD}}^{\text{weight}}, \mathcal{F}$) и ClusterWICER($sim_{\cos}, \mathcal{F}$) лучшее решение получено алгоритмом ClusterWICER($sim_{\text{NGD}}^{\text{weight}}, \mathcal{F}$), а худшее – ClusterWICER($sim_{\cos}, \mathcal{F}$) (рис. 3 и 4). Оно, скорее всего, является недостатком взвешивания слов, где при взвешивании не учитывается семантика слов.

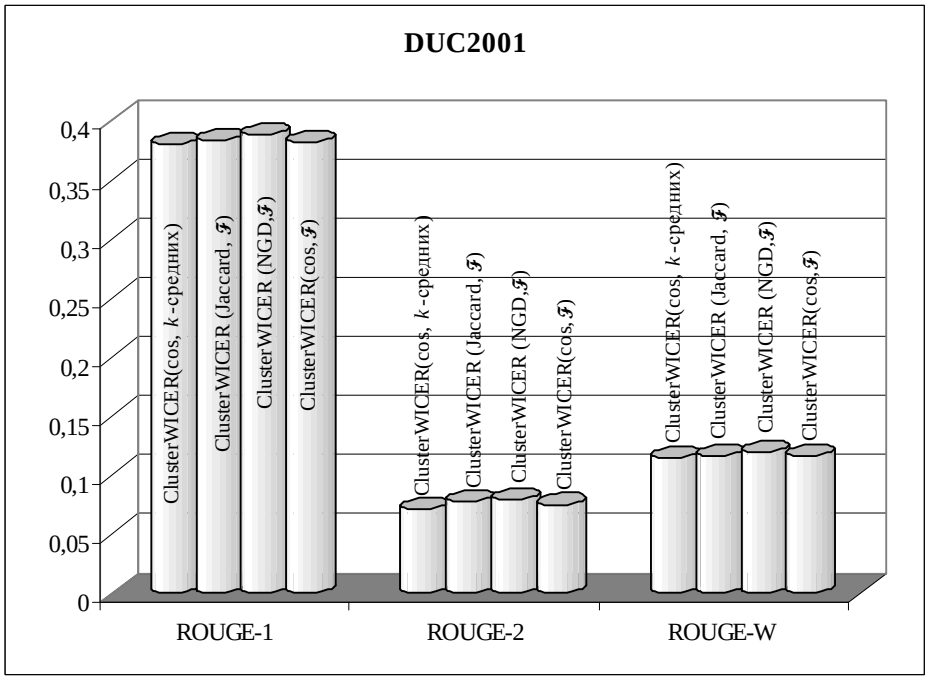

Рис.3. Влияние метода кластеризации и меры подобия на результат алгоритма ClusterWICER (DUC2001).

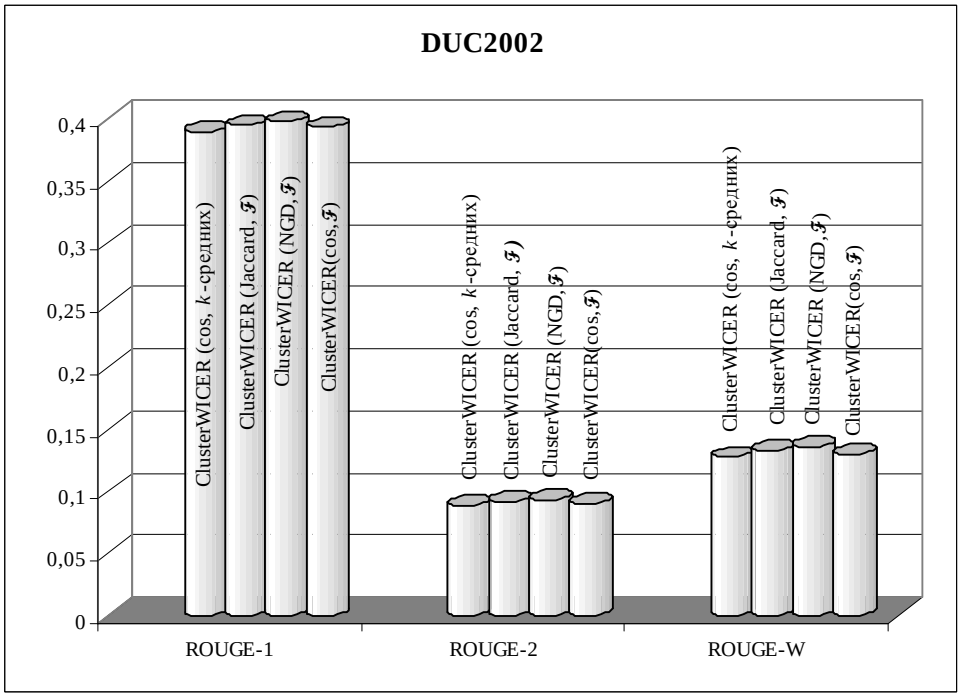

Рис.4. Влияние метода кластеризации и меры подобия на результат алгоритма ClusterWICER (DUC2002).

Заключение

Одна из проблем информационных перегрузок, с которыми мы сталкиваемся сегодня, заключается в том, что имеется много документов, относящихся к одной и той же теме. Метод многодокументного реферирования может помочь в решении данной проблемы. Многодокументное реферирование – это процесс автоматического создания сжатой версии совокупности документов, дающей пользователю полезную информацию. Предложен подход общего многодокументного реферирования, состоящий из двух этапов. На первом этапе с целью выявления тематических разделов осуществлена кластеризация предложений, а на втором этапе для определения степени информативности предложений произведено их ранжирование. Показано, что результаты реферирования зависят от методов кластеризации, алгоритма ранжирования, а также меры подобия. Результаты экспериментов показали, что выбор соответствующих методов кластеризации, алгоритма ранжирования и меры подобия в значительной степени влияет на точность методов реферирования.

Литература

1. Harabagiu S., Hickl A., Lacatusu V. Satisfying information needs with multi-document summaries // Information Processing and Management. 2007. V.43. №6. P.1619–1642.
2. Jones K. Automatic summarizing: the state of the art // Information Processing and Management. 2007. V.43. №6. P.1449–1481.
3. Moens M-F., Angheluta R., Dumortier J. Generic technologies for single- and multi-document summarization // Information Processing and Management. 2005. V.41. №3. P.569–586.
4. Zajic D., Dorr B.J., Lin J., Schwartz R. Multi-candidate reduction: sentence compression as a tool for document summarization tasks // Information Processing and Management. 2007. V.43. №6. P.1549–1570.
5. Zhang Y., Zincir-Heywood N., Milios E. World Wide Web site summarization // International Journal of Web Intelligence and Agents Systems. 2004. V.2. №1. P.39–53.
6. Antiqueira L., Oliveira O., Costa L., Nunes M. A complex network approach to text summarization // Information Sciences. 2009. V.179. №5. P.584–599.
7. Diao Q., Shan J. A new web page summarization method / Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR’06). Washington. USA. 2006. P.639–640.
8. Erkan G., Radev D. Lexrank: graph-based centrality as salience in text summarization // Journal of Artificial Intelligence Research. 2004. V.22. P.457–479.
9. Otterbacher J., Erkan G., Radev D. Biased LexRank: passage retrieval using random walks with question-based priors // Information Processing and Management. 2009. V.45. №1. P.42–54.
10. Zhang J., Xu H., Cheng X. GSPSummary: a graph-based sub-topic partition algorithm for summarization / Proceedings of the 2008 Asia Information Retrieval Symposium. Harbin. China. 2008. P.321–334.

11. Liu Y., Wang X., Zhang J., Xu H. Personalized PageRank based multi-document summarization / Proceedings of the First IEEE International Workshop on Semantic Computing and Systems (WSCS2008). Huangshan. China. 2008. P.169–173.
12. Zhang J., Cheng X., Wu G., Xu H. AdaSum: an adaptive model for summarization / Proceedings of the ACM 17th Conference on Information and Knowledge Management (CIKM'08). Napa Valley. USA. 2008. P.901–909.
13. Yeh J-Y., Ke H-R., Yang W-P. iSpreadRank: ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network // Expert Systems with Applications. 2008. V.35. №3. P.1451–1462.
14. Diligenti M., Gori M., Maggini M. A unified probabilistic framework for web page scoring systems // IEEE Transactions on Knowledge and Data Engineering. 2004. V.16. №1. P.4–16.
15. Wan X., Yang J., Xiao J. Manifold-ranking based topic-focused multi-document summarization / Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-2007). Hyderabad. India. 2007. P.2903–2908.
16. Тарасов С.Д. Алгоритм ранжирования связанных структур для задачи автоматического составления обзорных рефератов новостных сюжетов / Труды 11-й национальной конференции по искусственному интеллекту с международным участием (КИИ-2008). Дубна. Россия. 2008. Т.2. С.204–211.
17. Wan X., Yang J. Multi-document summarization using cluster-based link analysis / Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08). Singapore. 2008. P.299–306.
18. Aliguliyev R.M. A new sentence similarity measure and sentence based extractive technique for automatic text summarization // Expert Systems with Applications. 2009. V.36. №4. P.7764–7772.
19. Aliguliyev R.M. Clustering techniques and discrete particle swarm optimization algorithm for multi-document summarization // Computational Intelligence. 2009. V.25. №4.
20. Strehl A., Ghosh J. Value-based customer grouping from large retail data-sets / Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery. Orlando. USA. 2000. V.4057. P.33–42.
21. Padmanabhan D., Desikan P., Srivastava J. WICER: a weighted inter-cluster edge ranking for clustered graphs / Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'2005). Compiegne. France. 2005. P.522–528.
22. Lin C-Y. ROUGE: a package for automatic evaluation summaries / Proceedings of the Workshop on Text Summarization Branches Out. Barcelona. Spain. 2004. P.74–81.
23. http://duc.nist.gov

UOT 004.02:004.032.26
Alıquliyev R.M.
AMEA İnformasiyaTexnologiyaları İnstitutu, Bakı, Azərbaycan
aramiz@iit.ab.az a.ramiz@science.az

Cümlələri klasterləşdirmə və ranqlama yolu ilə sənədlər çoxluğunun referatlaşdırılması

Sənədlər çoxluğunun icmal referatının yaradılması üçün yanaşma təklif olunmuşdur. Təklif olunan yanaşma iki mərhələdən ibarətdir. Birinci mərhələdə sənədlər çoxluğunun bəhs etdiyi mövzular aşkarlanır, ikinci mərhələdə isə informativ cümlələr seçilir. Mövzular cümlələr çoxluğunu klasterləşdirmə yolu ilə, informativ cümlələr isə ranqlama alqoritminin köməyilə müəyyən olunur. Eksperiment nəticəsində məlum oldu ki, təklif olunan klasterləşdirmə metodları k-ortalar metodundan, ranqlama alqoritmi isə məşhur PageRank və HITS alqoritmlərindən yaxşı nəticə göstərir.
Açar sözlər: *icmal referat, klasterləşdirmə metodu, ranqlama alqoritmi.*

Aliguliyev R.M.
Institute of Information Technology ANAS, Baku, Azerbaijan
aramiz@iit.ab.az a.ramiz@science.az

Multidocument summarization through clustering and ranking of sentences

Two staged unsupervised approach to multidocument summarization is offered. At the first stage a set of documents are grouped into topics and at the second stage the informative sentences are extracted. The topics are defined through sentences clustering, and the informative sentences are extracted using the ranking algorithm. It is shown that summarization results depend on the clustering method, ranking algorithm and similarity measure. Experiments on open benchmarks DUC2001 and DUC2002 have shown that the offered clustering methods and ranking algorithm outperform the k-means method and the ranking algorithms PageRank and HITS.
Key words: *multidocument summarization, sentence clustering, sentence ranking.*