Симанков В. С., Толкачев Д. М.

РАЗРАБОТКА ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ СИСТЕМЫ ПОЛУЧЕНИЯ РЕЛЕВАНТНЫХ ДАННЫХ И ЗНАНИЙ В СЕТИ ИНТЕРНЕТ

Аннотация: *Статья посвящена разработке методических положений и алгоритмов получения релевантных данных и знаний в сети Интернет. Под релевантными данными и знаниями понимается информация, необходимая для решения какой-либо задачи или проблемы. В статье исследуются вопросы, связанные со смысловым сжатием информации, обеспечением семантической связности текстов, определением смыслового подобия текстов или фраз, а также с автоматическим поиском кратких и точных ответов на вопросы. Исследования учитывают особенности сети Интернет как источника огромного объёма неструктурированной информации. При проведении исследования использовались системный подход, теория алгоритмов, алгебра логики, теория множеств и сравнительный анализ. Представлен общий алгоритм проблемно-ориентированного автореферирования. Освещены вопросы поиска семантических связей между предложениями. Приведены методики составления интегрированного автореферата и выявления смыслового подобия двух текстов. Разработан алгоритм поиска ответов на вопрос. Представлены результаты разработки информационно-аналитической системы получения релевантных данных и знаний в сети Интернет.*
Ключевые слова: *данные, знания, Интернет, поисковые системы, проблемно-ориентированное автореферирование, семантические связи, местоимённые анафоры, регулярные выражения, смысловое подобие, тернарное выражение*

Для эффективного принятия любых управленческих решений необходимо наличие достаточного объёма данных и знаний, касающихся решаемой проблемы. Лицо, принимающее решения, может не обладать всеми необходимыми сведениями, поэтому в качестве источника актуальной информации часто используют сеть Интернет, чья роль в современном обществе неуклонно возрастает.

На основании проведённых ранее исследований сформируем общий алгоритм получения релевантных данных и знаний в сети Интернет (рисунок 1).

DOI: 10.7256/2305-6061.2014.3.13396

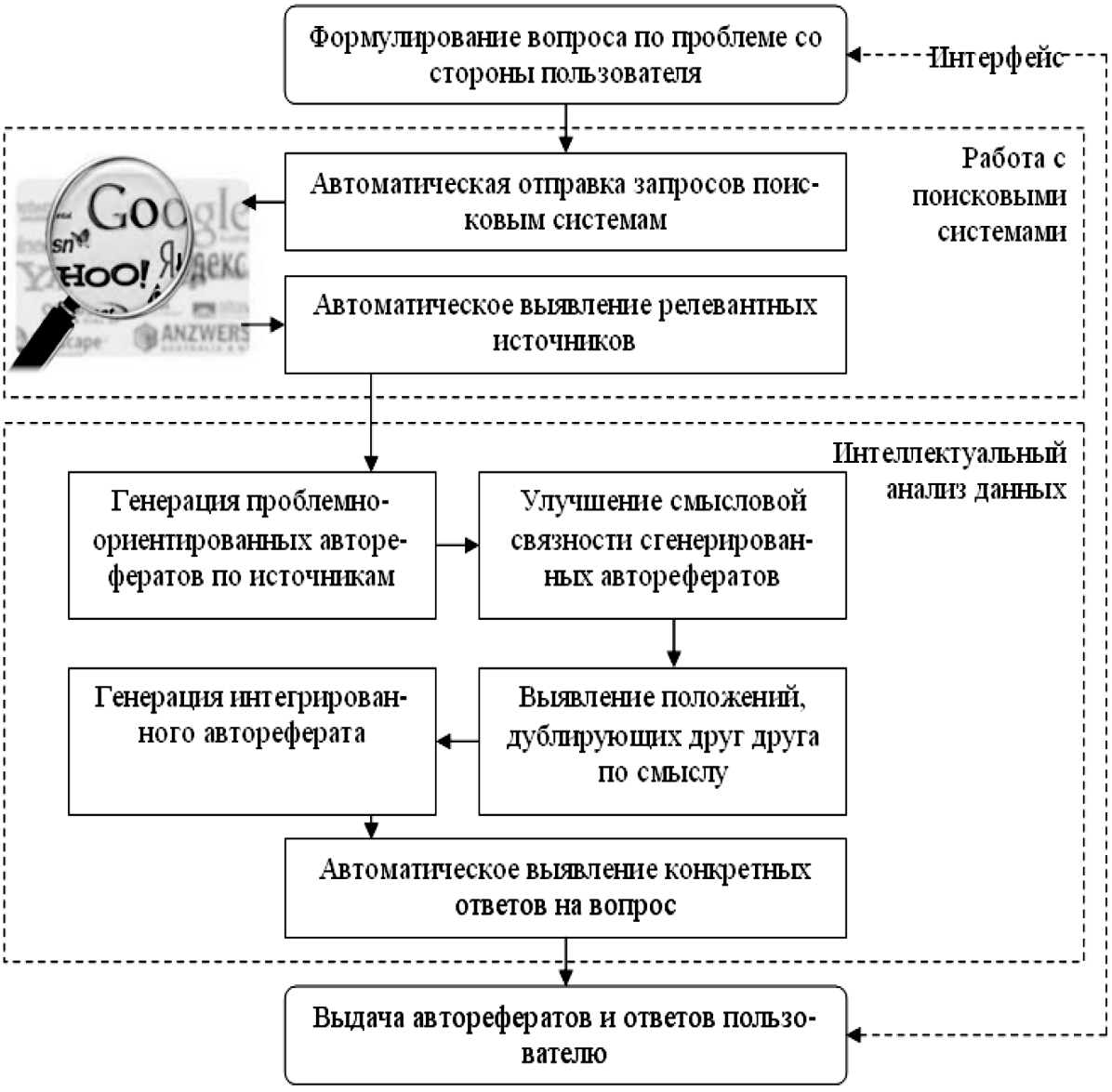

Рисунок 1 – Общий алгоритм поиска данных и знаний в Интернет

В настоящее время существует ряд подходов, которые позволяют получить краткую информационную выжимку из веб-источника. К ним относятся сниппеты, ассоциативный поиск и автореферирование.

Для создания наиболее универсального средства получения данных по проблеме и ответов на вопрос из сети Интернет была предложена комбинация классического автореферирования, ассоциативного поиска и сниппетов – проблемно-ориентированное автореферирование (ПОА) [1].

Основу ПОА составляет индикаторный метод квазиреферирования. Схематично алгоритм ПОА представлен на рисунке 2.

DOI: 10.7256/2305-6061.2014.3.13396

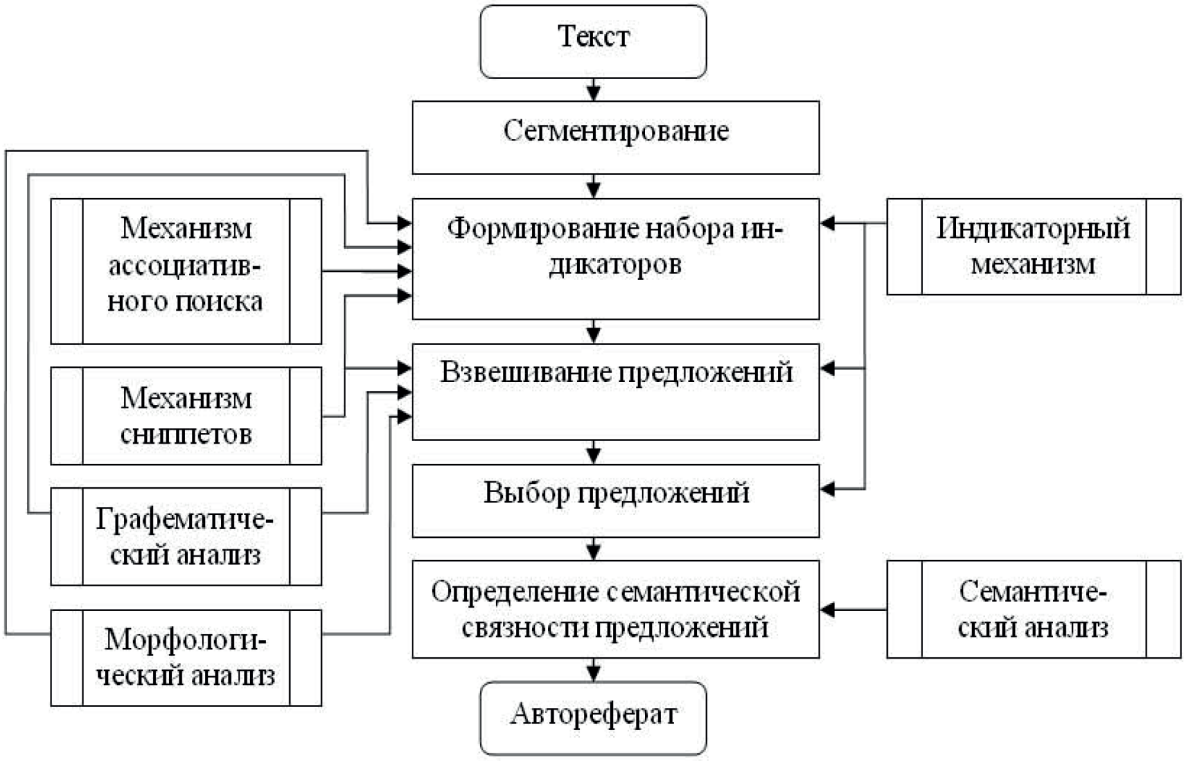

Рисунок 2 – Алгоритм ПОА

В процессе сегментирования происходит разбивка текста на абзацы и предложения. Индикаторы формируются из четырёх основных групп:

- слова (за исключением союзов, предлогов, частиц и междометий) из запроса (описания проблемы) и их морфологические формы (QUES);
- слова и словосочетания из универсального словаря «действий» (ACT); такой словарь должен содержать слова и их формы, которые с существенной долей вероятности указывают на то, что в предложении говорится о каких-либо действиях, путях решения проблемы или выводах, например: необходимо, следует, выполнить, решается, таким образом, сделать и т.д.;
- синонимы слов из описания проблемы и их формы (ASSOC); для этого нужно наличие словаря синонимов; данная группа, в отличие от первых двух, может быть пустой;
- слова из тематического словаря (TOPIC), т.е. словаря, составленного специально для определённой области знаний и содержащего наиболее важные слова и термины, характерные для данной области; наличие такого словаря повысит эффективность проблемно-ориентированного автореферирования при заполнении базы знаний по определённой теме; эта группа также может быть пустой.

DOI: 10.7256/2305-6061.2014.3.13396

Взвешивание предложений осуществляется по формуле:

$$\mathrm{ws} = \begin{cases} 0, CH = 0 \\ w_Q \cdot \sum_{i=1}^{n}(ind_i) + w_A \cdot \sum_{j=1}^{m}(A_j \cdot ind_j) + w_{AS} \cdot \sum_{k=1}^{o}(ind_k) + w_T \cdot \sum_{l=1}^{p}(T_l \cdot ind_l), CH = 1 \end{cases}, \quad (1)$$

где ws – вес предложения, w_Q – вес группы индикаторов QUES, w_A – вес группы индикаторов ACT, w_{AS} – вес группы индикаторов ASSOC, w_T – вес группы индикаторов TOPIC, n – число индикаторов в QUES, m – число индикаторов в ACT, o – число индикаторов в ASSOC, p – число индикаторов в TOPIC, A_j – вес j-ого индикатора в ACT, T_l – вес l-ого индикатора в TOPIC, ind_i – количество появлений i-ого индикатора в предложении; CH – логическая функция проверки адекватности предложения.

Количество появлений отдельного индикатора можно определить с помощью алгоритма, представленного на рисунке 3.

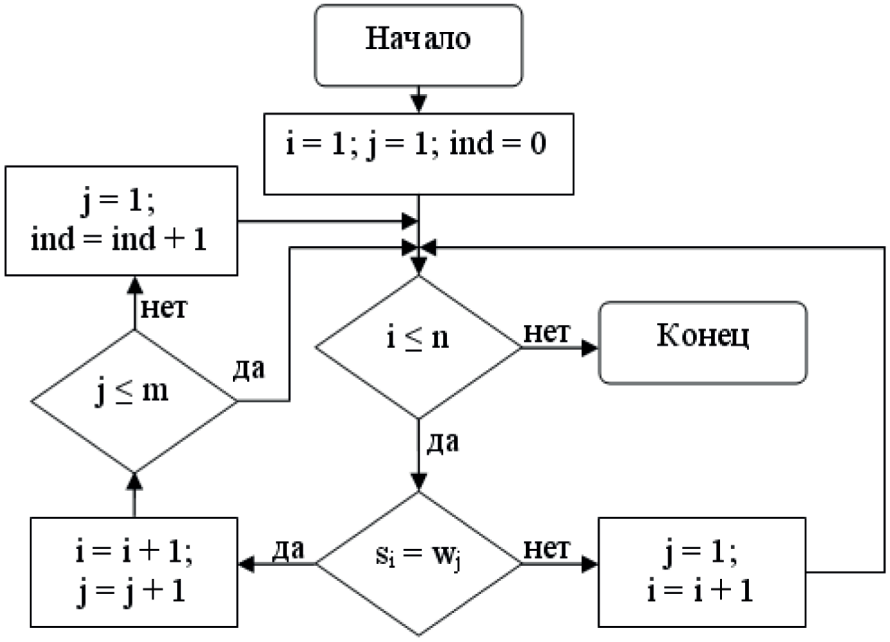

Рисунок 3 – Алгоритм определения количества появлений индикатора в предложении

На рисунке 3: n – число символов в предложении, m – число символов в индикаторе, ind – количество появлений индикатора, s_i – i-ый символ предложения, w_j – j-ый символ индикатора.

Введение логической функции CH соответствует механизму сниппетов, который предполагает отбрасывание информационного шума. В неё можно включить три правила:

- предложение должно содержать больше пяти символов;
- предложение должно содержать больше одного слова;
- предложение должно содержать хотя бы один глагол.

DOI: 10.7256/2305-6061.2014.3.13396

Формализуем СН:

$$CH = [(ns > 5) \wedge (nw > 1) \wedge (\exists w_i \in S : \exists end = \{s_j, s_{j+1}, \ldots, s_{j+k}\} \in w_i : (k \geq 0) \wedge$$
$$\wedge (\neg \exists s_l \in w_i : l > (j + k)) \wedge (end \in END) \wedge (w_i \notin EXC))] \quad , \qquad (2)$$

где: ns – число символов в предложении; nw – число слов в предложении; w_i – i-ое слово; S – предложение; s_j – j-ый символ предложения; END – множество глагольных окончаний; EXC – множество распространённых слов-исключений, имеющих глагольные окончания, но не являющихся глаголами.

Выбор определённого числа предложений из текста для включения их в автореферат происходит так:

$$AUTO = \{S \mid S \in TEXT\}, \forall S_i \in AUTO \neg \exists S_j \notin AUTO : (S_j \in TEXT) \wedge (ws_j > ws_i) \quad , (3)$$

где $AUTO$ – автореферат, $TEXT$ – исходный текст.

Число предложений, составляющих автореферат, – $|AUTO|$ – должно быть настраиваемым: как в виде абсолютной величины, так и в процентах к размеру анализируемого текста.

Для определения семантической связности предложений существуют некоторые методы. Так, достаточно эффективный алгоритм поиска местоимённых анафор предложен в [2]. При автореферировании любого типа требуется определять следующие семантические связи: местоимённую анафору, организацию логических связей и вводные слова [3]. Для их обнаружения было предложено использование набора эвристических правил в виде регулярных выражений – шаблонов, состоящих из символов и метасимволов и задающих правила поиска.

Будем использовать регулярные выражения PCRE (Perl Compatible Regular Expressions – перл совместимые регулярные выражения). Основы их синтаксиса приведёны в [4]. Рассмотрим некоторые его элементы:

/ – начало шаблона; ^ – начало строки; () – выделяют подшаблон; [] – строковой класс, или набор символов, которые могут быть в данном месте; [^] –набор символов, которые не могут быть в данном месте; * – любое количество символов; | – операция «ИЛИ»; . – любой символ, кроме разрыва строки; \s – любой символ пробела; \ – экранирование спецсимволов; /iu – конец регистронезависимого шаблона; ? – ноль или один символ.

Механизм разбора регулярных выражений PCRE основывается на использовании недетерминированных конечных автоматов [5], [6]. Существуют готовые программные решения, которые могут быть использованы для реализации этого механизма.

Приведём пример регулярного выражения для выявления местоимённой анафоры «он». В соответствии с правилами русского языка, можно выделить следующие эвристики для её выявления: зависимое предложение должно содержать местоимение «он» прописными или строчными буквами в любом месте, и до него не должно быть запятых и точек с запятой; запятые и точки с запятой в прямой речи не учитываются. Регулярное выражение примет вид:

/^([^,;]*|.*[,!?.]\s[\-––]\s[^,;]*)\sон[\s.,:;!?]/iu (4)

DOI: 10.7256/2305-6061.2014.3.13396

В (4) предполагается, что в начале предложения есть один пробел.

Общий алгоритм поиска семантических связей при автореферировании изображён рисунке 4.

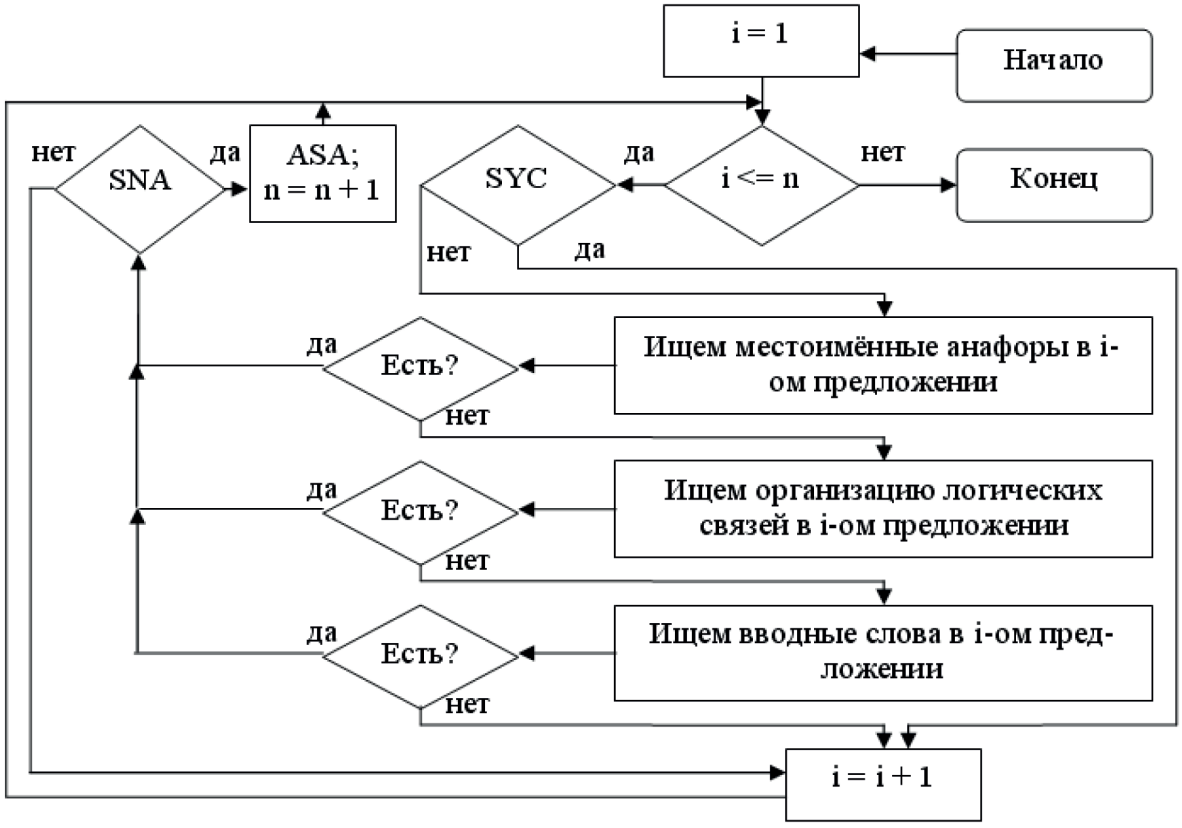

Рисунок 4 – Алгоритм поиска семантических связей при автореферировании

На рисунке 4: n – число предложений в автореферате; SYC – i-ое предложение уже проверялось?; SNA – предложения из основного текста, которое идёт перед i-ым, нет в автореферате?; ASA – добавляем предложение из основного текста, которое идёт перед i-ым, в автореферат, делаем его i-ым, а к номерам всех последующих добавляем единицу.

Проблемно-ориентированные авторефераты, полученные описанным выше методом, будут включать наиболее важные сведения, содержащиеся в проанализированных веб-источниках и касающиеся определённого вопроса.

Для агрегации информации из проблемно-ориентированных авторефератов различных веб-страниц, т.е. для составления дайджеста, или интегрированного автореферата, можно предложить достаточно простой метод. Его суть заключается в том, что в дайджест попадают предложения из различных источников, обладающие максимальным весом, сформированным на этапе построения ПОА. В результате интегрированный автореферат будет содержать только наиболее важные положения из всех исходных авторефератов.

Возможны и другие варианты. Если в дайджесте требуется в той или иной степени

DOI: 10.7256/2305-6061.2014.3.13396

осветить все источники, он будет составляться как совокупность укороченных версий исходных авторефератов. При этом «укороченные версии» формируются в полной аналогии с формированием «полных», разница заключается только в более высоком коэффициенте сжатия.

Важным вопросом при агрегации информации из различных источников является определение смыслового подобия текстов и отдельных положений в них. Смысловое подобие двух текстов можно определить как сходность содержащейся в них информации. Были выделены критерии смыслового подобия текстов авторефератов и методика его определения [7]. В соответствии с ней, для определения смыслового подобия двух ПОА вначале необходимо при помощи морфологического анализа или с помощью стеммера Портера выделить основы слов bw.

По словарю синонимов/гипонимов каждая основа приводится к базовому варианту b, если она отлична от него. При этом у частиц «не» выделение основы не проводится, вместо этого они удаляются, а следующие за ними слова получают значение коэффициента koef = –1, тогда как его начальное значение для всех слов koef = 1.

Каждая основа b_i получает вес w_i, который зависит от присутствия bw_i в словарях индикаторов и вычисляется по формуле:

$$w_i = 1 + w_Q \cdot Q_i + w_A \cdot A_i + w_{AS} \cdot AS_i + w_T \cdot T_i \quad , \qquad (5)$$

где w_Q – вес группы индикаторов QUES; Q_i – определяет, входит ли bw_i в QUES, и принимает значение 1, если входит, иначе – 0; w_A – вес группы индикаторов ACT; A_i – определяет, входит ли bw_i в ACT, и принимает значение веса индикатора bw_i в ACT, если входит, иначе – 0; w_{AS} – вес группы индикаторов ASSOC; AS_i – определяет, входит ли bw_i в ASSOC, и принимает значение 1, если входит, иначе – 0; w_T – вес группы индикаторов TOPIC; T_i – определяет, входит ли bw_i в TOPIC, и принимает значение веса индикатора bw_i в TOPIC, если входит, иначе – 0.

Далее необходимо из множества базовых основ B составить множество уникальных базовых основ UB:

$$UB = \{ub \mid (ub \in B) \wedge (\neg \exists i \neq j : ub_i = ub_j)\} \quad , \qquad (6)$$

При этом вес ub вычисляется как среднее арифметическое весов b, которые объединил ub:

$$wa_i = \frac{\sum w_j}{|\{b_j\}|},\ j \in J : \forall j1, j2 \in J \exists (b_{j1} = b_{j2} = ub_i) \quad , \qquad (7)$$

где wa_i – вес ub_i.

Коэффициент подобия двух авторефератов вычисляется по формуле:

DOI: 10.7256/2305-6061.2014.3.13396

$$K_{sim(p,q)} = \frac{2 \cdot \sum_{i=1}^{n\cap} \left(\min(wap_i, waq_i) \cdot \min(fp_i, fq_i) \cdot \left[\frac{\sum_{j=1}^{fp_i} koefp_j}{fp_i} = \frac{\sum_{j=1}^{fq_i} koefq_j}{fq_i} \right] \right)}{\sum_{i=1}^{np} wa_i \cdot fp_i + \sum_{i=1}^{nq} wa_i \cdot fq_i}, \quad (8)$$

где: $K_{sim(p,q)}$ – коэффициент подобия авторефератов p и q; $n\cap$ – количество общих уникальных базовых основ слов ub у авторефератов p и q; $\min(fp_i, fq_i)$ – функция определения минимального значения из количества $b_j = ub_i$ в автореферате p и количества $b_j = ub_i$ в автореферате q соответственно; np, nq – количество ub у авторефератов p и q соответственно; $\min(wap_i, waq_i)$ – функция определения минимального значения из веса ub_i в автореферате p и веса ub_i в автореферате q соответственно; $koefp_j$, $koefq_j$ – значения коэффициентов $koef$ для b_j авторефератов p и q соответственно.

В формуле (8) учитывается положительный и отрицательный характер высказывания, и полученный с помощью неё коэффициент подобия будет стремиться к единице лишь при пересечении понятийного состава, совпадении важных элементов текста и отсутствии противоположности заложенных в текстах идей одновременно.

Для определения общих и различных положений авторефераты разбиваются на абзацы P в соответствии с таким разбиением исходного текста. Допустимо осуществлять разбиение на отдельные предложения. Коэффициент их подобия вычисляется так:

$$K_{sim(P1,P2)} = \begin{cases} \dfrac{2 \cdot \sum_{i=1}^{n\cap} \min\left(\dfrac{\sum_{j=1}^{fP1_i} w_j}{fP1_i}, \dfrac{\sum_{j=1}^{fP2_i} w_j}{fP2_i}\right) \cdot \min(fP1_i, fP2_i)}{\sum_{i=1}^{nP1} \sum_{j=1}^{fP1_i} w_j + \sum_{i=1}^{nP2} \sum_{j=1}^{fP2_i} w_j}, NOT = 1 \forall i \in [1, n\cap] \\ 0, \exists i \in [1, n\cap] : NOT = 0 \end{cases}, \quad (9)$$

где $K_{sim(P1,P2)}$ – коэффициент подобия абзацев $P1$ и $P2$; NOT – логическая функция (10); все прочие обозначения аналогичны соответствующим в (8).

$$NOT = \left[\frac{\sum_{j=1}^{fP1_i} koefP1_j}{fP1_i} = \frac{\sum_{j=1}^{fP2_i} koefP2_j}{fP2_i} \right], \quad (10)$$

Если $K_{sim(P1,P2)}$ выше некоторого установленного на основе экспериментальных данных порогового значения, то считается, что в абзацах или предложениях говорится об одном и том же, и их можно не дублировать. Таким образом, перебирая различные

DOI: 10.7256/2305-6061.2014.3.13396

пары абзацев и оставляя только один из имеющих высокое значение $K_{sim(P1,P2)}$, можно получить обобщённый автореферат по нескольким источникам. Он также может быть подвергнут процедуре автореферирования, если источников достаточно много.

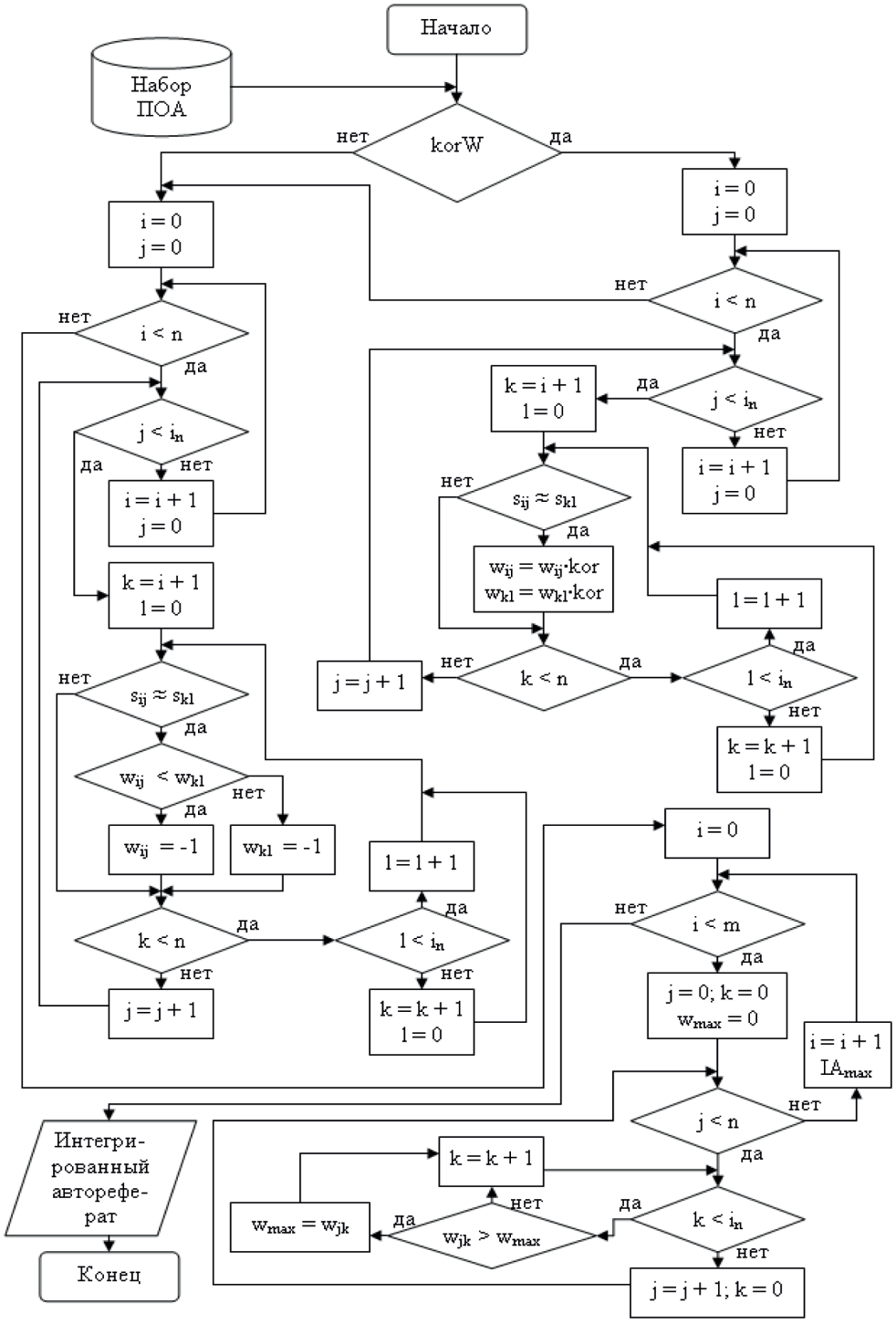

Рисунок 5 – Алгоритм построения интегрированного автореферата

DOI: 10.7256/2305-6061.2014.3.13396

На рисунке 5: korW – выполняется ли корректировка весов; n – число ПОА; $\mathrm{i_n}$ – число предложений в i-ом ПОА; $\mathrm{s_{ij}} \approx \mathrm{s_{kl}}$ – j-ое предложение i-ого ПОА определено как дублирующее по смыслу l-ого предложения k-ого ПОА; $\mathrm{w_{ij}}$ – вес j-ого предложения i-ого ПОА; kor – значение коэффициента, в соответствии с которым осуществляется корректировка весов; m – число предложений в интегрированном автореферате; $\mathrm{IA_{max}}$ – предложение с весом $\mathrm{w_{max}}$ включается в интегрированный автореферат, вес этого предложения становится равен -1.

Кроме применения оценки смыслового подобия отдельных положений текста для исключения дублирующих друг друга из интегрированного автореферата, возможно следующее её использование. Если какое-либо предложение одного из исходных авторефератов получило высокий коэффициент подобия с другим предложением другого автореферата, то вес такого предложения увеличивается, и оно в первую очередь попадает в дайджест. Подобный метод учитывает тот факт, что если одна и та же мысль повторяется в разных источниках, есть существенная вероятность того, что эта мысль важнее и достоверней прочих. Данный аспект отражён в алгоритме построения интегрированного автореферата на рисунке 5 при выполнении корректировки весов.

Автоматическое получение прямых и точных ответов на вопросы пользователя является актуальным направлением развития информационных технологий. Для решения этой задачи в качестве источника данных и знаний целесообразно использовать сеть Интернет. Тогда вопрос становится эквивалентен запросу к поисковой системе в сети Интернет, а ответы будут содержаться в найденных веб-источниках.

Для обработки вопроса и генерации ответа из текстового массива обычно используются четыре вида анализа: графематический, морфологический, синтаксический и семантический [8].

Был предложен подход к поиску ответов на вопросы, в той или иной степени использующий все основные виды анализа и основанный на принципах работы системы START [9]. Относительно видов анализа его можно представить так [10]:

- Графематический анализ – выделение слов и устойчивых словосочетаний.
- Морфологический анализ – определение характеристик слов и выделение словарных основ.
- Синтаксический анализ – сопоставление структуры вопросительного предложения со структурой ответа. Использование шаблонов совместно с результатами морфологического анализа (тернарные выражения + S-правила + Лексикон).
- Семантический анализ – учёт синонимичных и гипонимических замен (WordNet).

Общий алгоритм поиска ответов на вопрос изображён на рисунке 6.

DOI: 10.7256/2305-6061.2014.3.13396

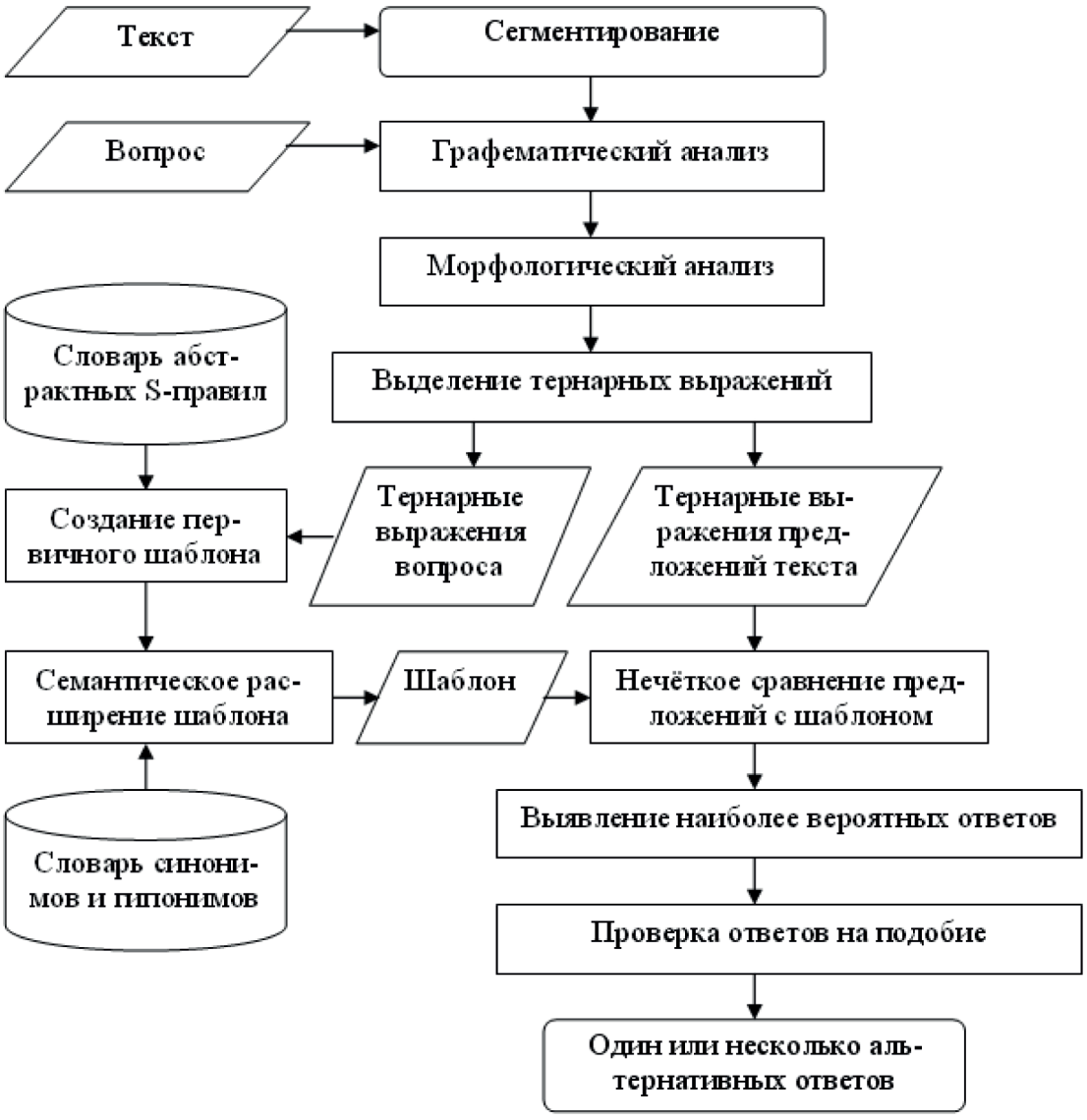

Рисунок 6 – Алгоритм поиска ответов на вопрос

В результате мы получим важные сведения по вопросу из отдельных источников, выделим из них и обобщим наиболее значимую информацию и сформулируем краткие ответы на поставленный вопрос.

Для практической реализации всех этапов представленного на рисунке 1 алгоритма нами был спроектирован и разработан прототип информационно-аналитической системы получения релевантных данных и знаний в сети Интернет (ИАС «IntellS»). Прототип имел небольшой словарь и обрабатывал лишь некоторые типы вопросов. Приведём сравнительный пример работы ИАС «IntellS» с двумя ближайшими аналогами – отечественной русскоязычной системой AskNet [11] и зарубежной англоязычной системой START (таблица 1).

DOI: 10.7256/2305-6061.2014.3.13396

Таблица 1. Сравнительный пример работы систем

Вопрос	AskNet	START	ИАС «IntellS»
Где проходили первые олимпийские игры? (Where were the first Olympic games?)	1) в городе Шамони 2) в австрийском Инсбруке 3) городом в до	I think you can find the relevant information here: http://en.wikipedia.org/wiki/1896_Summer_Olympics (Я думаю, вы можете найти релевантную информацию здесь: http://en.wikipedia.org/wiki/1896_Summer_Olympics)	1) Принято считать, что первые игры состоялись в 776 году до новой эры и были организованы в честь бога Зевса в почитаемом греками святилище Олимпия, расположенном в западной части Пелопоннесского полуострова. 2) Первые Всемирные соревнования, аналогичные древнегреческим Олимпийским играм, прошли в 1896 году в Афинах. 3) С1924 г. кроме Олимпийских игр, которые проходят летом, стали устраиваться и зимние Игры, чтобы могли состязаться и лыжники, конькобежцы и другие спортсмены, которые занимаются зимними видами спорта. 4) Первые Зимние олимпийские игры состоялись в 1924. Поначалу зимние и летние Игры проходили в один и тот же год, но начиная с 1994, они проводятся в разное время.

Системе START вопрос задавался на английском языке, тогда как AskNet и ИАС «IntellS» – на русском. Соответственно, и ответ система START дала на английском, на русском же ответили AskNet и ИАС «IntellS».

В качестве поисковой системы, которой отправлялись запросы на формирования перечня веб-источников, у ИАС «IntellS» использовался Яндекс [12]. При этом ответы искались по представленному выше алгоритму из набора проблемно-ориентированных авторефератов.

Из таблицы видно, что ИАС «IntellS» отвечает на вопросы более полно и точно, чем аналогичные системы. Таким образом, предложенные методики и алгоритмы обладают практической эффективностью.

В результате проведённого исследования можно сделать следующие выводы:

- На основе индикаторных методов квазиреферирования с использованием графе-

DOI: 10.7256/2305-6061.2014.3.13396

матического, морфологического и семантического анализов, а также механизмов ассоциативного поиска и сниппетов разработан общий алгоритм проблемно-ориентированного автореферирования.

- На основании сформулированных методических положений разработан алгоритм поиска семантических связей между предложениями при автореферировании.
- Разработаны методы составления интегрированных авторефератов из нескольких источников на основании проведённого аналитического обзора.
- С учётом необходимости выявления противоречий и определения сходства отдельных предложений и абзацев при проблемно-ориентированном автореферировании разработана методика автоматической оценки смыслового подобия текстов.
- С использованием тернарных выражений, абстрактных S-правил и нечётким сравнением предложений с шаблоном разработан алгоритм поиска ответов на вопрос в сети Интернет.
- На основе предложенных методических положений и алгоритмов разработан прототип информационно-аналитической системы получения релевантных данных и знаний в сети Интернет, обладающий практической эффективностью.

Библиография :

1. Симанков В.С., Толкачев Д.М. Проблемно-ориентированное автореферирование как инструмент поиска данных и знаний // Наука вчера, сегодня, завтра / Сб. ст. по материалам XIV междунар. науч.-практ. конф. № 7 (14). Новосибирск: Изд. «СибАК», 2014. – с. 31-35.
2. В.Е. Абрамов, Н.Н. Абрамова, Е.В. Некрасова, Г.Н. Росс. Статистический анализ связности текстов по общественно-политической тематике. Труды 13й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2011, Воронеж, Россия, 2011. – с. 127-133.
3. Симанков В.С., Толкачев Д.М. Обеспечение смысловой связности текста автореферата // Научная дискуссия: инновации в современном мире. № 7 (27): сборник статей по материалам XXVII международной заочной научно-практической конференции. – М., Изд. «Международный центр науки и образования», 2014. – с. 12-16.
4. Perl regular expressions [Электронный ресурс]. Режим доступа: http://perldoc.perl.org/perlre.html (22.10.2014).
5. Фридл Дж. Регулярные выражения, 3-е издание. – Пер. с англ. – СПб.: Символ-Плюс, 2008. – 608 с.
6. Oliver Müller. Pattern Matching with Regular Expressions in C++ [Электронный ресурс]. Режим доступа: http://www.tldp.org/LDP/LGNET/issue27/mueller.html (22.10.2014).
7. Симанков В.С., Толкачев Д.М. Автоматическая оценка смыслового подобия текстов // Технические науки – от теории к практике / Сб. ст. по материалам XXXVII междунар. науч.-практ. конф. № 8 (33). Новосибирск: Изд. «СибАК», 2014. – с. 26-33.

DOI: 10.7256/2305-6061.2014.3.13396

8. К.Х. Ким, А.П. Савинов. Синтаксический анализатор для вопросно-ответной системы. Известия Томского политехнического университета, Т. 315. № 5, 2009. – с. 133-138.
9. START, Natural Language Question Answering System [Электронный ресурс]. Режим доступа: http://start.csail.mit.edu/index.php (22.10.2014).
10. Симанков В.С., Толкачев Д.М. Поиск ответов на вопросы в сети Интернет // Инновации в науке / Сб. ст. по материалам XXXVI междунар. науч.-практ. конф. № 8 (33). Новосибирск: Изд. «СибАК», 2014. – с. 28-35.
11. Семантическая поисковая система AskNet [Электронный ресурс]. Режим доступа: http://www.asknet.ru/ (22.10.2014).
12. Яндекс [Электронный ресурс]. Режим доступа: http://www.yandex.ru/ (22.10.2014)

References:

1. Simankov V.S., Tolkachev D.M. Problemno-orientirovannoe avtoreferirovanie kak instrument poiska dannykh i znanii // Nauka vchera, segodnya, zavtra / Sb. st. po materialam XIV mezhdunar. nauch.-prakt. konf. № 7 (14). Novosibirsk: Izd. «SibAK», 2014. – s. 31-35.
2. V.E. Abramov, N.N. Abramova, E.V. Nekrasova, G.N. Ross. Statisticheskii analiz svyaznosti tekstov po obshchestvenno-politicheskoi tematike. Trudy 13i Vserossiiskoi nauchnoi konferentsii «Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii» – RCDL'2011, Voronezh, Rossiya, 2011. – s. 127-133.
3. Simankov V.S., Tolkachev D.M. Obespechenie smyslovoi svyaznosti teksta avtoreferata // Nauchnaya diskussiya: innovatsii v sovremennom mire. № 7 (27): sbornik statei po materialam XKhVII mezhdunarodnoi zaochnoi nauchno-prakticheskoi konferentsii. – M., Izd. «Mezhdunarodnyi tsentr nauki i obrazovaniya», 2014. – s. 12-16.
4. Perl regular expressions [Elektronnyi resurs]. Rezhim dostupa: http://perldoc.perl.org/perlre.html (22.10.2014).
5. Fridl Dzh. Regulyarnye vyrazheniya, 3-e izdanie. – Per. s angl. – SPb.: Simvol-Plyus, 2008. – 608 s.
6. Oliver Müller. Pattern Matching with Regular Expressions in C++ [Elektronnyi resurs]. Rezhim dostupa: http://www.tldp.org/LDP/LGNET/issue27/mueller.html (22.10.2014).
7. Simankov V.S., Tolkachev D.M. Avtomaticheskaya otsenka smyslovogo podobiya tekstov // Tekhnicheskie nauki – ot teorii k praktike / Sb. st. po materialam XXXVII mezhdunar. nauch.-prakt. konf. № 8 (33). Novosibirsk: Izd. «SibAK», 2014. – s. 26-33.
8. K.Kh. Kim, A.P. Savinov. Sintaksicheskii analizator dlya voprosno-otvetnoi sistemy. Izvestiya Tomskogo politekhnicheskogo universiteta, T. 315. № 5, 2009. – s. 133-138.
9. START, Natural Language Question Answering System [Elektronnyi resurs]. Rezhim dostupa: http://start.csail.mit.edu/index.php (22.10.2014).
10. Simankov V.S., Tolkachev D.M. Poisk otvetov na voprosy v seti Internet // Innovatsii v nauke / Sb. st. po materialam XKhXVI mezhdnar. nauch.-prakt. konf. № 8 (33). Novosibirsk: Izd. «SibAK», 2014. – s. 28-35.
11. Semanticheskaya poiskovaya sistema AskNet [Elektronnyi resurs]. Rezhim dostupa: http://www.asknet.ru/ (22.10.2014).
12. Yandeks [Elektronnyi resurs]. Rezhim dostupa: http://www.yandex.ru/ (22.10.2014)

DOI: 10.7256/2305-6061.2014.3.13396