Т.В. Шманина

ИНФОРМАЦИОННАЯ СИСТЕМА ДЛЯ ПОДДЕРЖКИ ПРОЦЕССА ПРОВЕДЕНИЯ ИССЛЕДОВАНИЙ НА ОСНОВЕ ЛИТЕРАТУРНЫХ ИСТОЧНИКОВ

ВВЕДЕНИЕ

Над решением технологических и исследовательских задач работает огромное число коллективов по всему миру, непрерывно производя большие объемы полезного знания. В ходе исследований коллективами могут применяться два основных метода: выбор оптимального решения текущей задачи из множества существующих решений (либо его построение на основе последних) и создание принципиально нового, оригинального решения проблемы.

Зачастую наиболее рациональным способом решения поставленных задач является первый путь. Однако он требует наличия информации об имеющихся разработках, которая чаще всего доступна в виде научных публикаций и технической литературы, объемы которой чрезвычайно велики и продолжают расти. Кроме того, с ростом объема производимого знания возрастает степень специализации исследователей, что не позволяет им быть осведомленными о многих потенциально полезных для их практики методах, подходах, а возможно, и целых областях знания.

Возникает проблема создания автоматических или полуавтоматических средств, упрощающих процесс поиска потенциальных решений поставленной перед исследователем задачи. Наилучшие потенциальные решения задач могут быть обнаружены в не связанных напрямую с тематикой решаемой задачи областях знаний. Поэтому такие методы поиска информации, как обычный (возможно, семантический) информационный поиск, кластеризация документов с целью группировки их по основной тематике и некоторые другие, зачастую лишь в незначительной степени упрощают задачу поиска решения, так как остается необходимость поиска взаимосвязей между методами, подходами и т.п.

Возможным ответом в вопросе поиска потенциальных решений некоторой задачи может стать автоматизация (полная или частичная) метода исследования на основе литературных источников, предложенного Доном Свенсоном в 1986 году.

В процессе исследования на основе литературных источников новое знание не генерируется на основе экспериментальных данных или путем логического вывода. Вместо этого производится попытка найти взаимосвязь между существующими знаниями, уже полученными экспериментальным или дедуктивным путем, посредством выделения ранее незамеченных взаимосвязей между сущностями, описанными в литературных источниках. Эта техника широко используется на практике уже долгое время [2].

Выделение таких скрытых взаимосвязей получило название «связывание Свенсона». Чаще всего этот процесс рассматривался исследователями в контексте биомедицинских задач. Типичным примером связывания Свенсона является следующий. Предположим, исследователь занимается разработкой лекарства от заболевания A (например, синдрома Рейно) и известно, что заболевание вызывается веществом B. Кроме того, известно, что препарат C уменьшает концентрацию вещества B в организме, и, таким образом, C может быть лекарством от болезни A. Однако последние данные публиковались отдельно от литературы, посвященной лечению A, поэтому взаимосвязь между болезнью A и препаратом C могла быть упущена (см. рис. 1). Связывание Свенсона имеет своей целью выделение таких взаимосвязей и предоставление их исследователю в явном виде.

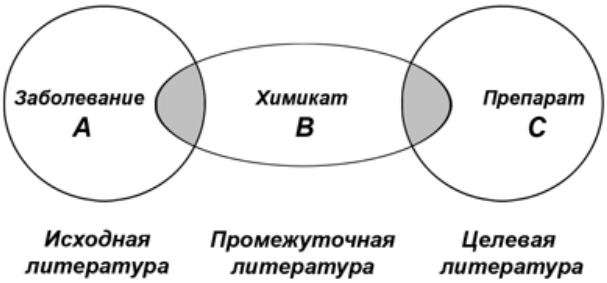

Рис. 1. Связывание Свенсона для заболевания А и препарата С

В ряде работ была предпринята попытка создания инструментов, автоматизирующих процесс построения связываний Свенсона для нахождения потенциальных решений задач непосредственно в виде концептов. Боль-

шинство этих работ было ориентировано на биомедицинскую область и основывалось на идее поиска взаимосвязей между разрозненными литературными источниками по общим ключевым понятиям, содержащимся в текстах. Упомянутые подходы, в отличие от предлагаемого в данной работе, стремятся выделить точные термины, определяющие потенциальное решение задачи, и предоставить их пользователю в виде упорядоченного по релевантности списка. К числу инструментов, применяющихся при создании таких моделей, относятся методы информационного поиска, латентное семантическое индексирование, использование внешних тезаурусов, онтологий и различных статистических характеристик для ранжирования потенциальных решений и выявления взаимосвязи между терминами [2].

Целью проводимого автором исследования является разработка подхода, позволяющего осуществлять анализ конечной локальной коллекции текстовых документов с целью построения сети взаимосвязи тем и проблем, затрагиваемых в литературе, для упрощения процесса исследования на основе имеющихся литературных источников.

МЕТОД АВТОМАТИЗАЦИИ ПРОЦЕССА ИССДЕДОВАНИЯ НА ОСНОВЕ ЛИТЕРАТУРНЫХ ИСТОЧНИКОВ

В целях автоматизации процесса исследования на основе литературных источников автор данной работы предлагает метод, стремящийся имитировать процесс исследования научной и технической литературы человеком.

Принцип действия предлагаемого метода следующий. Алгоритм получает на вход формулировку некоторой проблемы в виде одного или нескольких ключевых терминов, а также конечную локальную коллекцию документов научного или технического плана. В процессе работы алгоритм сначала выделяет в коллекции документов темы, непосредственно связанные с введенной проблемой, а затем итеративно выделяет в еще не просмотренной литературе темы, связанные с темами, выделенными на предыдущей итерации, пытаясь таким образом имитировать поведение исследователя. В результате алгоритм строит ориентированный граф взаимосвязи основных тем, затронутых в литературе (Рис. 2). В этом графе ориентированные маршруты, ведущие от вершины, соответствующей исходной проблеме, реализуют цепочку шагов, предпринимаемых исследователем при последовательном поиске и изучении литературы, касающейся вопросов, связанных напрямую или косвенно с введенной проблемой. В качестве результата работы алгоритма пользователю будет предоставляться построен-

ный граф взаимосвязи тем, наличие которого в готовом виде может помочь ускорить подбор необходимой для исследования литературы.

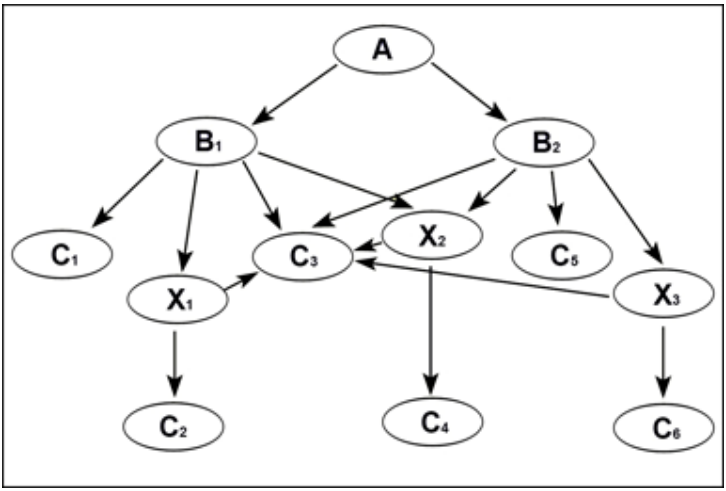

Рис. 1 Граф взаимосвязи тем, построенный в процессе поиска решения проблемы A

В основе описанного метода лежит подход, предложенный Р. Костоффом в 2003 году [1]. Однако, в силу ряда особенностей последнего, таких как предполагаемый полуавтоматический режим работы и ограничение на глубину строящегося графа тем, было принято решение произвести модификацию данного подхода.

Алгоритм построения графа взаимосвязи тем

Определения:

Коллекция документов D – набор неструктурированных текстов произвольной тематики. Предлагаемый в данной работе подход рассчитан на работу с локальной конечной коллекцией документов научного или технического плана на английском языке.

Ключевая фраза – слово или словосочетание w, извлеченное из текста $d \in D$ и имеющее высокую содержательную значимость для текста. В наилучшем случае является термином, отражающим одну из тем, содержащихся в тексте.

Тема T **–** набор семантически тесно связанных между собой ключевых фраз, $T = \{w_1, \ldots, w_N\}$. В наилучшем случае все ключевые фразы, образующие тему T, относятся к одной и той же достаточно узкой тематике, и поэтому определенная таким образом тема должна быть способна дать исследователю четкое представление о тематическом содержании документа.

Проблема P – с точки зрения пользователя представляет собой текстовую формулировку задачи, связанные тематики для которой необходимо найти. Задается в виде словосочетания или набора словосочетаний на естественном языке. С точки зрения алгоритма проблема есть тема, составленная из нескольких ключевых фраз.

Граф взаимосвязи тем – ориентированный помеченный граф $G = (V, E)$, где V – множество вершин графа, и каждая вершина $v \in V$ помечена множеством ключевых фраз, образующих некоторую тему, построенную алгоритмом, а $E \subset V \times V$ – множество ребер, такое, что $\forall v_1, v_2 \in V : e = (v_1, v_2) \in E \leftrightarrow$ тема, которой помечена вершина v_2, была получена алгоритмом при поиске литературы, непосредственно касающейся темы v_1.

Уровень темы T в графе G – номер итерации алгоритма построения графа тем, на котором была сгенерирована тема *T*.

Множество просмотренных документов $Prohib(D)$ – множество документов коллекции, на которых в ходе работы алгоритма уже производилось выделение тем. Документы из этого множества не принимают участия в дальнейшем процессе построении множества тем (то есть документы «изымаются» из коллекции после первого просмотра). Данная мера позволяет реализовать поддержку идеи связывания Свенсона: система стремится выделить связи между темами, описанными в разрозненной, непересекающейся литературе. Для таких тем вероятность иметь неизвестные исследователю взаимные связи выше.

Множество запрещенных ключевых фраз при построении тем уровня k $Prohib(k)$ – набор ключевых фраз, которые не должны присутствовать в документах, непосредственно связанных с некоторой темой *T* уровня *k*–1. Во множество запрещенных ключевых фраз при построении тем уровня *k* попадают все ключевые фразы, составляющие темы уровней 0, …, $k - 2$. Таким образом, система ограничивает возможность выделения тех документов, в которых упоминаются темы, по которым уже производился поиск. Последнее требуется методологией проведения исследований на основе литературных источников: алгоритм имеет целью выделить те

связи между проблемой и темами, которые ранее еще не были построены (если и не в исследовательской практике, то, по крайней мере, самим алгоритмом). Поэтому темы, выделенные в процессе построения графа взаимосвязи тем на более ранних итерациях, считаются известными и далее не представляют интереса. Кроме того, посредством применения этой меры производится попытка избежать построения чрезмерно разветвленного и зашумленного итогового графа тем.

Дополнительно каждой теме сопоставляется набор документов, из которых были выделены составляющие ее ключевые фразы, что позволяет пользователю осуществлять навигацию по коллекции документов при изучении построенного графа.

Предлагаемый алгоритм:

Вход: формулировка проблемы в виде набора фраз на естественном языке и локальная коллекция документов научного или технического плана.

Выход: граф взаимосвязи тем.

Алгоритм:

Шаг 0. Формулировка проблемы в терминах алгоритма.

Набору фраз, введенному пользователем, сопоставляется тема T_{01} уровня 0.

Шаг *k*. Построение множества тем $\{T_{k,1},\ldots,T_{k,m_k}\}$ уровня *k* по множеству тем $\{T_{k-1,1},\ldots,T_{k-1,m_{k-1}}\}$ уровня *k*–1.

- Для каждой темы $T_{k-1,i}$ уровня *k*–1 выполнить:
  - Поиск литературы $D_{k,i}=\left\{d_{k,i,1},\ldots,d_{k,i,j_i}\right\}\subset D\setminus Prohib(D)$, непосредственно связанной с темой $T_{k-1,i}$ (процесс поиска подробно описан далее). Если $D_{k,i}=\varnothing$, завершаем данную итерацию и переходим к теме $T_{k-1,i+1}$.
  - Выделение множества тем $[T]_{k,i}=\{T_{k,i_1},\ldots,T_{k,m_{ki}}\}$, затронутых в найденной литературе $D_{k,i}$ (метод выделения тем подробно описан далее).
- Поместить множество документов, просмотренных на данной итерации, во множество всех просмотренных документов: $Prohib\left(D\right)=Prohib(D)\cup\cup_i D_{k,i}$.

- Поместить множество ключевых фраз, составляющих темы уровня k–1, во множество запрещенных ключевых фраз: $Prohib(k+1) = Prohib(k) \cup \cup_i T_{k-1,i}$.
- Произвести перекомпоновку тем уровня k для получения множества тем $\{T_{k,1},\ldots,T_{k,m_k}\}$:
  - Если $\{[T]_{k,1},\ldots,[T]_{k,m_{k-1}}\} = \varnothing$, итеративный процесс выделения тем завершается на шаге $k = N$. Иначе:
  - Для каждой темы $T_{k,n} \in \cup_i [T]_{k,i}$ поочередно пытаемся объединить ее с другими темами $T_{k,m} \in \cup_i [T]_{k,i}$ (критерий объединения приведен далее). Таким образом, устраняются дубликаты тем, которые могли быть построены для некоторых $T_{k-1,n}$ и $T_{k-1,m}$, $n \neq m$ (внутри же любой $[T]_{k,i}$ темы заведомо не пересекаются по построению). Получаем множество тем $\{\tilde{T}_{k,1},\ldots,\tilde{T}_{k,m_k}\}$.
  - Для всех пар различных тем $\tilde{T}_{k,n},\tilde{T}_{k,m} \in \{\tilde{T}_{k,1},\ldots,\tilde{T}_{k,m_k}\}$ устраняем их пересечение (метод описан далее). Получаем множество тем $\{T_{k,1},\ldots,T_{k,m_k}\}$.

Шаг N+1. Перекомпоновка графа тем.

Для каждой пары соседних уровней k-1 и k графа тем, для каждой темы $T_{k-1,i} \in \{T_{k-1,1},\ldots,T_{k-1,m_{k-1}}\}$ и $T_{k,j} \in \{T_{k,1},\ldots,T_{k,m_k}\}$ определяется необходимость объединения этих тем, и производится объединение в случае такой необходимости (критерий объединения тем приведен далее).

На этом построение графа тем завершается.

Приведенный алгоритм конечен в силу того, что

1. Коллекция текстовых документов, на основе которой строится граф, конечна.
2. Каждый раз после завершения процесса выделения тем на шаге k вся найденная и обработанная на k-й итерации литература помещается в «запрещенное» множество, по которому поиск на дальнейших итерациях не производится. Таким образом, рано или поздно коллекция документов будет исчерпана.
3. Алгоритм чаще всего заканчивает работу до момента исчерпания коллекции документов в силу пополнения множества запрещенных

ключевых фраз и особенности формирования запроса на поиск связанной литературы (см. следующий раздел).

Поиск литературы по заданной теме

Поиск литературы по заданной теме производится стандартными поисковыми средствами. В частности в данной работе для поиска документов используется поисковый сервер Apache Solr, реализованный на базе библиотеки полнотекстового поиска Apache Lucene [11].

Поиск производится следующим образом. Имея некоторую тему на *k*-й итерации алгоритма, для которой необходимо найти связанную литературу, система генерирует запрос к поисковой системе по принципу:

1. В запрос включаются все ключевые слова, образующие тему.
2. К запросу добавляются все термины, образующие темы, построенные на итерациях 0 – *k*–2, с запрещающим поисковым оператором. Иначе говоря, эти фразы либо совсем не должны содержаться в найденных документах, либо ранг таких документов в общем списке понижается.

Например, в случае системы Solr, которая поддерживает стандартный синтаксис поисковых запросов для ИПС общего назначения (таких как Google), запрос Q, сгенерированный по теме $B = \{b_1, \ldots, b_N\}$, при условии, что множество ключевых фраз, образующих темы на итерациях 0 – *k*–2, равно $D = \{d_1, \ldots, d_M\}$, будет иметь вид:

$$Q = "b_1\ b_2 \ldots b_N\ \ -d_1 -d_2 \ldots -d_M\ ".$$

Из множества документов, полученных от поисковой системы, выбираются только те документы, оценка релевантности запросу которых, вычисленная поисковой системой, превышает некоторый порог. Данный порог релевантности должен регулироваться пользователем и, вообще говоря, может зависеть от объема и состава текстовой коллекции, по которой производится поиск.

Выделение тем для множества документов

Для того чтобы по выделенному набору документов сформировать набор затрагиваемых в них тем, необходимо выполнить следующие шаги:

1. Выделить ключевые фразы из каждого документа.
2. Произвести кластеризацию на полученном наборе ключевых фраз с целью группировки семантически связанных ключевых фраз.

Выделение ключевых фраз. Библиотека Maui

Качество сформированных тем в значительной степени зависит от качества ключевых фраз, выделенных из текста и подаваемых на вход алгоритму кластеризации. В случае если важные термины и ключевые фразы будут упущены, темы, построенные по множеству выделенных фраз, будут иметь неполную структуру, либо некоторые из них вовсе могут быть упущены. Если будет выделено слишком много общеупотребительных фраз, возникнет опасность генерации искусственных кластеров, которые будут сформированы за счет перекрывания множеств общеупотребительных фраз из разных текстов. Кроме того, в этом случае могут быть построены ошибочные связи между полученными темами за счет таких «общеупотребительных» связок. Такое явление может послужить причиной необоснованного разрастания конечного графа тем по количеству вершин и ребер и снижению «содержательности» тем, соответствующих вершинам.

Поэтому для выделения ключевых фраз из текстов было решено использовать Maui – библиотеку для автоматического индексирования научной литературы. В основе работы Maui лежит использование статистических методов и алгоритмов машинного обучения для выделения наиболее значимых ключевых фраз из текстовых документов. Maui обладает высокими показателями качества извлечения терминов из документов, сравнимыми с показателями качества индексирования текстов людьми-индексаторами. Подробно с ее устройством и принципом работы можно познакомиться в работах [3] и [4].

Кластеризация

Для формирования тем, иными словами, для группировки тесно связанных ключевых слов, используется метод кластеризации [7]. В общем случае, кластеризация – процесс разбиения заданного множества точек на группы на основе атрибутов этих точек.

Существует несколько классов алгоритмов кластеризации, для которых понятие кластера варьируется и, следовательно, кластеры, построенные разными алгоритмами, значительно отличаются по свойствам. Автором работы был произведен обзор основных классов кластеризационных методов, с результатами которого можно ознакомиться в [5].

Учитывая полученные сведения об особенностях наиболее распространённых алгоритмов кластеризации, для решения поставленной задачи было решено применять центроидный метод кластеризации k-means++ в сочетании с заданием метрик для определения семантических расстояний между ключевыми фразами и заданием требуемого для формирования тем числа кластеров равным $\sqrt{n/2}$, где n – число кластеризуемых точек [8, 9]. При этом задаваемое число кластеров выражает интуитивное понятие об оптимальном выборе числа кластеров, который представляет собой баланс между максимальной компрессией данных путем помещения их в один кластер и максимальной точностью разбиения на кластеры, когда каждой точке множества соответствует свой кластер.

Семантическая дистанция между фразами

Для задания расстояний между точками при кластеризации предлагается использовать метрики на основе меры семантической дистанции между ключевыми фразами.

Существует два основных подхода к измерению семантической дистанции между фразами:

1. Основанный на использовании сторонних источников знаний, например, специализированных словарей, тезаурусов, онтологий.
1. Такие методы позволяют напрямую вычислить показатель близости двух слов или фраз и основаны на предположении, что используемый тезаурус содержит всю необходимую для расчетов информацию. Использование таких мер затруднено тем, что:
   a. не для всех языков построены тезаурусы;
   b. в случае, если тезаурус имеется, он может быть неполон.
2. С примерами тезаурусных мер семантической близости можно ознакомиться в [6].
3. Статистические методы.
4. Многими исследователями было показано, что семантику слова можно определить, пусть даже приближенно, через типичный контекст его употребления. Поэтому, а также в силу наличия богатого тренировоч-

ного материала, для реализации предлагаемого алгоритма было принято решение сосредоточиться на применении статистических методов.

5. Статистические методы опираются на на подсчет характеристик, основанных на совместной встречаемости слов. При этом все слова рассматриваются как точки в *N*-мерном пространстве, и задача определения семантической дистанции сводится к двум основным шагам – задание координат точек в пространстве и вычисление дистанции между точками. Последнее требует подбора подходящей метрики.

В целях уменьшения размерности решаемой методом кластеризации задачи *N*-мерное пространство, в котором производится представление и кластеризация точек, строится по набору T выделенных на данном шаге ключевых фраз следующим образом:

1. Набор T рассматривается в качестве фиксированного на данном шаге словаря V_C, $|V_C| = N$.
2. Все ключевые фразы $v \in V_C$ нумеруются.
3. Теперь каждое слово или фраза $w \in V$ (бесконечного словаря) может быть представлено в виде *N*-мерного вектора величин совместной встречаемости $\overrightarrow{f_w} = (f_1, \ldots, f_N)$, где f_i указывает, насколько часто слово w встречается совместно со словом v_i.

Таким образом, каждая ключевая фраза представляется в виде точки в *N*-мерном пространстве.

В качестве координат точек f_i необходимо брать величины, выражающие степень семантической близости фраз с фразами из V_C и обладающие свойством: величина тем больше, чем больше семантическая близость двух рассматриваемых слов или фраз. Для этого автором работы используется величина Pointwise Mutual Information (PMI). Для произвольных фраз v и w данная величина имеет вид:

$$PMI(v, w) = \frac{f_{v,w}}{f_v f_w},$$

где $f_{v,w}$ – частота совместной встречаемости фраз v и w в некотором корпусе текстов, f_v – частота фразы v в этом корпусе. В качестве корпуса текстов при этом берется подаваемая на вход алгоритму коллекция документов *D*. Слова v_i и w считаются встретившимися совместно, если v_i в некотором тексте встретилось на расстоянии, не превышающем *N* слов от

слова w Величина N должна задаваться пользователем. Для экспериментов N было взято равным 20.

Следует отметить, что именно PMI была выбрана для реализации предлагаемого метода в силу ее большей адекватности в качестве меры семантической близости по сравнению с распространенными бинарной и абсолютной частотами совместной встречаемости слов [6].

Метрики для вычисления семантической дистанции между ключевыми фразами

Введение метрики для вычисления расстояний между ключевыми фразами дает возможность производить кластеризацию слов на основе информации об их семантической близости.

Для измерения дистанции между двумя векторами, соответствующими ключевым фразам можно использовать различные известные метрики. Для изучения влияния выбора метрики на качество кластеров были выбраны 4 известные метрики:

1. Расстояние по Манхэттену (метрика, порождаемая нормой L_1):

$$dist_{L_1}(\vec{x},\vec{y}) = \sum_{i=1}^{N} |x_i - y_i|.$$

2. Евклидова метрика (метрика, порождаемая нормой L_2):

$$dist_{L_2}(\vec{x},\vec{y}) = \left(\sum_{i=1}^{N} (x_i - y_i)^2\right)^{\frac{1}{2}}.$$

3. Метрика на основе величины угла между нормированными векторами ключевых фраз:

$$dist_{cos}(\vec{x},\vec{y}) = \frac{\arccos(sim_{cos}(\vec{x},\vec{y}))}{\pi},$$

где величина sim_{cos} выражает степень семантической близости между ключевыми фразами на основе близости их векторов и равна косинусу угла между векторами $\vec{x}$ и $\vec{y}$:

$$sim_{cos}(\vec{x},\vec{y}) = \frac{\vec{x} \bullet \vec{y}}{\vec{x}_2 \vec{y}_2},$$

где $\vec{x} \bullet \vec{y} = \sum_{i=1}^{N} x_i \bullet y_i$ – скалярное произведение векторов $\vec{x}$ и $\vec{y}$.

4. Дивергенция Дженсена–Шеннона.
   В случае двух векторов дивергенция Дженсена–Шеннона задается следующей формулой:

$$dist_{JS}\left(\vec{x},\vec{y}\right)=\sqrt{\sum_i\left(x_i log_2\frac{2x_i}{x_i+y_i}+y_i log_2\frac{2y_i}{x_i+y_i}\right)}.$$

   Доказательство того, что функция $dist_{JS}\left(\vec{x},\vec{y}\right)$ действительно обладает всеми свойствами метрики, можно найти в [10].

Операции на множестве тем

Первые два этапа алгоритма – это перекомпоновка тем и перекомпоновка графа. Перекомпоновка тем производится с целью устранения дубликатов тем в конечном графе. Такие дубликаты могут возникать среди тем, построенных на одной итерации (во множествах тем, построенных по различным темам из предыдущей итерации), либо же среди тем, построенных на двух последовательных итерациях.

Определим две операции:

1. Объединение тем производится в случае, если мера близости двух тем превышает установленный порог $t, 0 \le t \le 1$.
2. Устранение пересечения тем производится в противном случае. При этом множество ключевых фраз, попавших в пересечение, относится к меньшей по мощности теме.

В качестве меры близости тем берется коэффициент Жаккара: если A и B – два множества, то

$$sim_J\left(A,B\right)=\frac{\left|A\bigcap B\right|}{\left|A\bigcup B\right|},\ 0\le sim_J\left(A,B\right)\le 1.$$

Пороговое значение t может варьироваться и устанавливается вручную. Для экспериментов величина порога объединения тем бралась из интервала $0.6 \le t \le 0.8$.

Область применения алгоритма

В силу ряда особенностей метода, предложенного автором данной работы, а именно, применения статистических методов для формирования тем, описывающих коллекцию документов, а также использования ключевых фраз, извлекаемых из документов, как главный характеризующий их признак, данный метод ориентирован на анализ научной и технической литера-

туры. Кроме того, очевидно, что успех применения данного метода будет зависеть от степени сформированности терминологии, используемой в анализируемой литературе.

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

Предложенный подход к автоматизации процесса исследования на основе литературных источников был реализован в прототипе информационной системы для поддержки процесса анализа литературных источников.

Разработанный прототип работает в трех режимах:

1. Режим обучения.
   Этот режим существует для тренировки Maui. При работе в данном режиме системе на вход подается набор текстов, в котором каждому тексту сопоставлен список ключевых слов. На основе поданной коллекции текстов система генерирует модель извлечения ключевых фраз для дальнейшего применения в режиме индексирования. Соответственно, если индексирование документов будет производиться для коллекции документов по фиксированной тематике, то для обучения рекомендуется брать коллекцию документов по этой же либо родственной ей тематике. Кроме того, обучение можно производить с использованием словаря предметной области, либо вообще без словаря.
2. Режим индексирования.
   В этом режиме система получает на вход коллекцию документов в произвольном текстовом формате, поддерживаемом системой. В данный момент система поддерживает такие форматы, как txt, html, xml, doc, pdf и некоторые другие. Также должна быть указана какая-либо из моделей извлечения, построенная на этапе обучения, и может указываться словарь, на основе которого необходимо извлекать терминологию из текстов (без словаря ключевые фразы извлекаются непосредственно из текста, в противном случае тексту приписываются термины из словаря). Для каждого поданного документа производится извлечение из него (или приписывание ему) ключевых фраз, которые сохраняются в базе данных. Таким образом, в процессе поиска извлечение ключевых фраз каждый раз не производится. Это ускоряет работу приложения ипозволяет использовать для извлечения ключевых фраз модель извлечения и, возможно, словарь, подхо-

дящие по тематике к индексируемой коллекции, что улучшает качество индексирования.

3. Режим поиска.
   В этом режиме системе на вход подается запрос от пользователя в виде набора фраз на естественном языке. Система производит построение графа взаимосвязи тем по алгоритму, изложенному в данной работе, на множестве проиндексированных документов (фактически, на множестве ключевых фраз, сохраненных в базе данных). Построенный граф выдается пользователю в текстовом формате.

Архитектура приложения

Реализованный прототип информационной системы написан на языке Java и имеет стандартную трехуровневую архитектуру: пользовательский интерфейс, алгоритмическую компоненту и интерфейс доступа к данным. Схематично архитектура приложения изображена на Рис. 2.

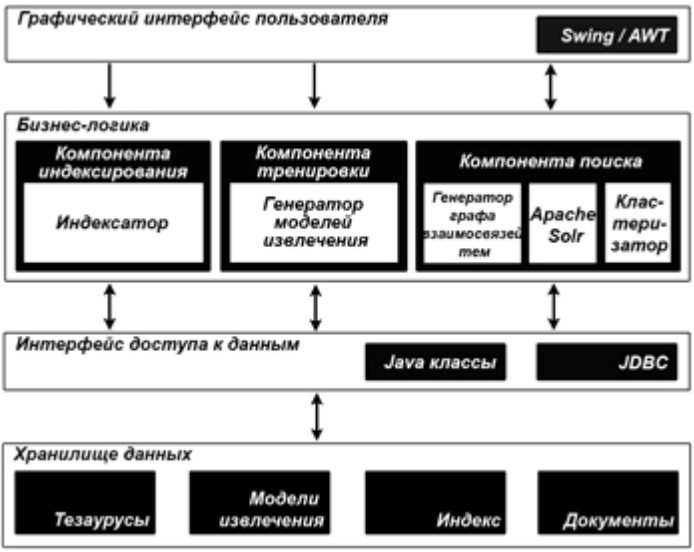

Рис. 2. Архитектура приложения

ТЕСТИРОВАНИЕ

Для тестирования приложения была взята коллекция, состоящая из 50 научных публикаций по тематике обработки сигналов и изображений (материалы конференции ICASSP 2002).

Предварительно приложение было обучено на множестве научных статей, состоящем из 800 документов по тематике обработки сигналов и изображений (материалы научной конференции ICASSP 2011). Обучающая выборка была подготовлена в полуавтоматическом режиме и представляла собой пары: публикация, преобразованная в текстовый формат, и файл, содержащий ключевые слова, которыми данная публикация была проиндексирована человеком-индексатором.

Тестирование производилось по 5 запросам. Варьировались следующие параметры:

- число ключевых фраз, сопоставляемых родительской теме при формировании списка дочерних тем (от него зависели число и размеры кластеров);
- метрика для алгоритма кластеризации;
- пороговое значение коэффициента Жаккара для операций над темами.

Для построенных графов оценивались следующие количественные параметры:

- общее число тем, выделенное по запросу из коллекции документов;
- число документов коллекции, вовлеченных в процесс построения графа тем;
- число итераций, проделанное алгоритмом при построении графа тем («глубина графа тем»).

Из результатов можно заключить, что при фиксированных метрике и пороге объединения тем и увеличении числа ключевых фраз, извлекаемых для каждой темы, глубина графа практически не варьируется, несмотря на разрастание графа в ширину, которое обусловлено прежде всего методом формирования кластеров. С другой стороны, разрастание графа связано также с увеличением общего числа вовлеченных в процесс поиска ключевых фраз, дающим большую возможность вариации тем, что подтверждается увеличением числа привлеченных к построению графа документов.

По результатам тестирования было также отмечено, что при фиксированном нижнем пороге числа ключевых фраз для получения более разветвленной структуры графа тем в случае метрики Дженсена-Шеннона необходимо принимать порог объединения тем равным 0.6, для метрики Манхет-

тен – 0.7, а для угловой метрики – 0.8. В случае метрики Евклида выбор порога неоднозначен, однако наибольшая вовлеченность текстовой коллекции была получена при величине порога объединения тем, равной 0.8.

Экспертом также была произведена качественная оценка информационной связности построенных графов тем. В целом было отмечено хорошее качество работы системы, а именно, информационная связность и содержательность построенных графов взаимосвязи тем с точки зрения потенциального исследователя. Была отмечена невысокая степень промахов системы при формировании тем – число фраз в кластерах, не подходящих к их основной тематике, было крайне незначительным. При этом сформированные темы-кластеры давали эксперту представление о подразумевающейся тематике и, таким образом, были содержательными с точки зрения человека-исследователя.

Наиболее эффективным сочетанием параметров по результатам качественной оценки работы системы являются:

- нижний порог числа ключевых фраз, равный 50;
- мера Дженсена-Шеннона;
- порог слияния тем по мере Жаккара, равный 0.6 (это значение дает наиболее разветвленный граф для метрики Дженсена-Шеннона, однако не влияет на информационную связность тем).

Также было проведено тестирование системы на большой выборке научных статей (1012 документов по тематике обработки сигналов и изображений – материалы конференции ICASSP 2002). Тестирование производилось с обозначенными выше оптимальными параметрами: нижний порог числа ключевых фраз, равный 50, порог слияния тем, равный 0.6, метрика Дженсена–Шеннона. Тестирование производилось по запросам «Compressing multi-component digital maps» и «Digital image watermarking», для которых при тестировании на малой коллекции были получены наихудшие качественные результаты в силу отсутствия в малой коллекции достаточной доли документов, связанных непосредственно или косвенно с тематикой запросов. Были получены следующие результаты:

1. Глубина построенных графов взаимосвязи тем была равна 8 в обоих случаях.
2. Оба графа взаимосвязи тем содержали порядка 2000 вершин.
3. Число вовлеченных файлов было равно 772 и 737 (из 1012), соответственно.

Заключение эксперта о качестве построенных графов взаимосвязи тем было следующим:

1. Для запроса «Compressing multi-component digital maps» были получены очень качественные результаты: очень удачная классификация ключевых фраз и документов по темам, адекватно отражающая структуру предметной области.
2. Результаты по второму запросу «Digital image watermarking» оказались менее очевидными. Полученная структура тем оказалась сложной и труднопонимаемой. Результаты работы системы на данном запросе требуют дополнительного пристального изучения экспертом, но также являются интересными.

ЗАКЛЮЧЕНИЕ

Итогом работы стала реализация прототипа информационной системы, в перспективе способной играть роль ассистента исследователя при поиске потенциальных решений исследовательских задач. В частности, разрабатываемая система имеет потенциал для выявления ранее не замеченных косвенных взаимосвязей между подходами, методами и технологиями, информация о которых была опубликована; система также способна выявлять и группировать литературу, которая может быть интересна исследователю при решении его задач.

Результаты работы, в частности, разработанный прототип информационной системы, могут служить основой для проведения экспериментальных исследований, касающихся проблемы автоматизации процесса исследования на основе литературных источников и дальнейшего развития методов автоматизации этого процесса.

Однако предложенный метод требует дальнейшей доработки. А именно, в дальнейшем могут быть исследованы вопросы:

- эффективности различных методов кластеризации в контексте решаемой задачи и методов улучшения качественных характеристик генерируемых тем;
- возможности и целесообразности комбинации статистических характеристик ключевых фраз с их характеристиками, вычисленными на основе тезаурусов и Википедии, при формировании тем;
- эффективности и целесообразности операций над темами в зависимости от их свойств (плотность тем, степень пересечения тем и т.д.);
- возможности и целесообразности разработки подходов для сокращения мощности генерируемого графа тем.

Таким образом, в будущем предполагается совершенствование предложенного метода и разработанного программного средства.

СПИСОК ЛИТЕРАТУРЫ

1. Kostoff R., Boylan R., Simons G. Disruptive technology roadmaps / Technol. Forecast. Soc. Change. – 71. – 2004. – P. 141–159 – Mode of access: http://www.cuaed.unam.mx/puel_cursos/cursos/d_gcfe_m_tres/modulo/modulo_3/m3-3.pdf
2. Ganiz M., Pottenger W., Janneck C. Recent Advances In Literature Based Discovery / Lehih University Technical Report LU-CSE-05-027. – 2005. – Mode of access: http://www.cse.lehigh.edu/~billp/pubs/JASISTLBD.pdf
3. Medelyan O., Frank E., Witten I. Human-competitive tagging using automatic keyphrase extraction / Proc. of the Internat. Conference of Empirical Methods in Natural Language Processing EMNLP-2009. – Singapore. – 2009. – 10 P. – Mode of access: http://www.cs.waikato.ac.nz/~olena/publications/emnlp2009_maui.pdf
4. Medelyan O. Human-competitive automatic topic indexing. PhD Thesis. – University of Waikato, New Zealand. – 2009. – 241 P. – Mode of access: http://www.cs.waikato.ac.nz/~olena/publications/olena_medelyan_phd_thesis_July2009.pdf
5. Шманина Т. В. Информационная система для поддержки процесса проведения исследований на основе литературных источников. Магистерская диссертация. – Новосибирский Государственный Университет, Россия. – 2012. – 64 С.
6. J. Hockenmaier Introduction to Natural Language Processing. Lectures at University of Illinois at Urbana-Champaign [Electronic resource]. – 2008. – Mode of access: http://www.cs.uiuc.edu/class/fa08/cs498jh/
7. Cluster analysis [Electronic resource]. – Mode of access: http://en.wikipedia.org/wiki/Cluster_analysis
8. D. Arthur, S. Vassilvitskii k-means++: the advantages of careful seeding / Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms SODA'07. – 2007. – pp. 1027-1035.
9. Determining the number of clusters in a data set [Electronic resource]. – Mode of access: http://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set
10. D. Endres, J. Schindelin A New Metric for Probability Distributions / IEEE Transactions on Information Theory. – 49 (7). – 2003.
11. Apache Solr Documentation [Electronic resource]. – Mode of access: http://lucene.apache.org/solr/