Р.Е. Суворов, И.В. Соченков

Определение связанности научно-технических документов на основе характеристики тематической значимости[1]

Аннотация. В статье представлены результаты исследования в области методов анализа неструктурированной текстовой информации. Разработан метод определения связанности научно-технических документов на основе характеристики тематической значимости. Выполнено экспериментальное сравнение результатов работы нескольких модификаций разработанных методов.

Ключевые слова: поиск тематически похожих документов, оценка тематического сходства, векторная модель, TF, IDF, характеристика тематической значимости, оценка методов информационного поиска, DCG.

Введение

На сегодняшний день в электронных библиотеках, базах данных и открытых источниках Интернета накоплено большое количество научно-технических документов в форме статей, авторефератов, диссертаций, монографий, отчетов, руководств, книг и учебников. В современных реалиях информационного общества существует потребность в быстром поиске и анализе этой информации для нужд исследователей, специалистов, экспертов. Для решения поисково-аналитических задач создан класс информационных систем, которые предоставляют пользователям такие функции по анализу крупных, постоянно пополняемых коллекций научно-технических документов, как:

– поиск и выборка информации, удовлетворяющей заданным критериям;

– систематизация информации – кластеризация и классификация на основе различных критериев для быстрого ознакомления пользователя с тематическим составом коллекции документов;

– анализ различных показателей коллекций с учетом их временной динамики и др.

Эти функции востребованы как исследователями и специалистами в предметных областях, так и экспертами в области наукометрии, изучающими структуру научного знания и научных сообществ.

Настоящее исследование посвящено задаче определения тематической связанности научно-технических документов. Решение этой задачи лежит в основе реализации таких аналитических функций поисково-аналитических систем, как:

– выявление научных направлений в предметных областях;

– выявление коллективов авторов, работающих в рамках одного научного направления;

– определение преемственности результатов исследований;

[1] Работа выполнена при финансовой поддержке Минобрнауки России по государственному контракту № 14.514.11.4024 в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы».

– поиск нечетких дубликатов документов;

– систематизация коллекций текстов путем кластеризации или классификации;

– формирование подборок документов, имеющих общую тематику с заданным документом-эталоном.

Автоматическое определение тематической связанности документов позволяет выполнить предварительную систематизацию информации: для заданного документа-эталона формируется множество тематически похожих документов. Это позволяет эксперту быстрее ознакомиться с тематикой документа-эталона и оценить состояние дел в предметной области путем анализа тематически похожих документов. Кроме того, для повышения качества и скорости решения таких задач, как автоматическое выявление коллективов авторов, определение преемственности исследований, целесообразно в качестве исходных данных рассматривать не всю коллекцию документов, а ее подмножество, состоящее из документов, тематически похожих на заданные эталонные документы.

Метод определения тематической связанности документов должен быть тесно интегрирован с другими методами обработки информации в поисково-аналитической системе. Несмотря на то, что существует значительное число методов тематической систематизации информации, многие из них не применимы в промышленных системах, не рассчитаны на большие объемы данных и не могут обеспечить результат за требуемое время.

Отметим, что в поисково-аналитических системах, как правило, предусмотрен автоматический классификатор в рамках предопределенной таксономии тематических рубрик. В качестве рубрик выступают, например, предметные области или отрасли науки [11]. Задача тематической классификации во многом близка к исследуемой, однако даже в наиболее подробной таксономии разбиение на классы ограничивается предметными областями, что не позволяет выделить группу документов, посвященную узкой проблематике в некоторой предметной области.

Целью настоящего исследования является разработка метода определения тематической связанности документов и экспериментальная проверка качества его работы. Также в ходе исследования изучен круг вопросов, касающихся практического применения этого метода в поисково-аналитической системе.

1. Методы определения тематической связанности документов

Рассмотрим коллекцию *c* текстовых документов. Множество документов, принадлежащих коллекции *c*, обозначим *D(c)*. Произвольно выбранный документ из этой коллекции назовем эталоном. Сформулируем задачу определения тематической связанности документов как формирование списка тематически похожих документов и его ранжирование по убыванию оценок тематического сходства с заданным эталонным документом.

Для представления информации о документах выбрана классическая векторная модель. Пространство признаков представляет собой множество лексических дескрипторов (ЛД). Под лексическим дескриптором понимаются [3, 5, 6]:

– отдельные лексемы как совокупности парадигматических форм (словоформ) одного слова, т.е. разные формы одного и того же слова не различаются;

– устойчивые словосочетания в канонических формах (главное слово приведено к словарной форме, а форма зависимых слов подчинена управлению главного слова) безотносительно различных форм.

Значениями признаков (координаты векторов, соответствующих документам) являются веса ЛД, определяющие информационную значимость ЛД в тексте документов. Оценку тематического сходства документов предлагается выполнять на основе сопоставления величин информационной значимости ЛД в текстах документов коллекции, используя ЛД с наибольшими весами. В соответствии с работой [6] рассмотрим следующие величины:

1. Инверсная частота ЛД *w* в произвольном множестве текстов τ.

$$IDF(w,\tau) = \log_2 \frac{|\tau|}{m(w,\tau)} \quad , \qquad (1)$$

где $m(w,\tau)$ – число документов в τ, содержащих лексический элемент *w*.

2. Вес ЛД в отдельном документе *d*:

$$lTF(w,d) = \log_{S(d)} \left(k(w,d) + 1 \right), \quad (2)$$

где $k(w,d)$ – количество вхождений ЛД *w* в тексте *T(d)* документа *d*, $S(d)$ – общее количество вхождений ЛД в текст документа *d*.

3. Как было отмечено выше, для документов коллекции *c* (или некоторого ее подмножества) может быть известна тематическая принадлежность документов в рамках априорно заданной таксономии. Тематика документов может быть определена вручную экспертами или автоматически системой с помощью процедуры классификации. В работах [6,10] для выявления наиболее информационно значимых ЛД предложено использовать изменение информативности ЛД *w* документа при отнесении его к некоторой тематической рубрике σ:

$$\Delta\tilde{I}(w,c,\sigma) = IDF_N(w, c \setminus D(\sigma)) - IDF_N(w, D(\sigma)), \quad (3)$$

$$\Delta\tilde{I}^+(w,c,\sigma) = \Delta\tilde{I}(w,c,\sigma) \cdot X(\Delta\tilde{I}(w,c,\sigma)), \quad (4)$$

где $X(z)$ – функция Хевисайда.

4. Информационная значимость ЛД *w* текста документа *d* в коллекции *c* определяется формулой:

$$v(w,d,c) = lTF(w,d)IDF(w,c). \quad (5)$$

5. Если известно, что эталонный документ относится к тематической рубрике σ, то значимость ЛД *w* текста документа *d* в коллекции *c* задается формулой:

$$\tilde{v}(w,d,c,\sigma) = lTF(w,d)\Delta\tilde{I}^+(w,c,\sigma). \quad (6)$$

Таким образом, каждому документу *d* соответствует вектор $V(d,c) = (v(w_1,d,c), v(w_2,d,c), ..., v(w_n,d,c))$, определяющий информационную значимость ЛД w_i, в тексте этого документа. Наравне с этим вектором весов рассматривается вектор $\tilde{V}(d,c,\sigma) = (\tilde{v}(w_1,d,c,\sigma), \tilde{v}(w_2,d,c,\sigma), ..., \tilde{v}(w_n,d,c,\sigma))$, определяющий информационную значимость ЛД w_i, в тексте документа *d* с учетом априорной информации о том, что *d* относится к тематической рубрике σ.

Выявление тематической связанности двух документов d_1, d_2 заключается в оценке тематического сходства, т.е. в расчете расстояния между векторами $\rho(V(d_1,c), V(d_2,c))$ (либо между векторами $\rho(\tilde{V}(d_1,c,\sigma), \tilde{V}(d_2,c,\sigma))$).

Для оценки тематического сходства двух документов использованы модификации следующих функций: косинусное расстояние [15], расстояние Хэмминга. Оценки тематического сходства и тематического различия (расстояния) связаны соотношением:

$$SIM(d_i,d_j) = 1 - \rho(d_i,d_j),$$

где $\rho(d_i,d_j)$ – оценка расстояния между документами d_i и d_j, $SIM(d_i,d_j)$ – оценка тематического сходства между документами d_i и d_j.

Будем считать, что два документа тематически связаны, если оценка тематического сходства превосходит некоторый заданный порог, величина которого подбирается на этапе настройки системы на основе анализа экспертных мнений о тематической похожести эталонных документов из обучающей выборки.

Оценки тематического сходства, основанные на косинусном расстоянии и расстоянии Хэмминга, являются неотрицательными, нормированными на единицу величинами и выражаются формулами:

$$SIM_{\cos}(d_i,d_j) = \frac{\sum_{w \in \hat{W}(d_i) \cap \hat{W}(d_j)} v(w,d_i,c)v(w,d_j,c)}{\sqrt{\sum_{w \in \hat{W}(d_i) \cap \hat{W}(d_j)} \left(v(w,d_i,c)\right)^2} \sqrt{\sum_{w \in \hat{W}(d_i) \cap \hat{W}(d_j)} \left(v(w,d_j,c)\right)^2}}, \quad (7)$$

$$SIM_{Ham}(d_i,d_j) = 1 - \frac{\sum_{w \in \hat{W}(d_i) \cap \hat{W}(d_j)} \left|v(w,d_i,c) - v(w,d_j,c)\right|}{\sum_{w \in \hat{W}(d_i) \cap \hat{W}(d_j)} \left(v(w,d_i,c) + v(w,d_j,c)\right)}, \quad (8)$$

где $SIM_{\cos}(d_i,d_j)$ - косинусная оценка тематического сходства; $SIM_{Ham}(d_i,d_j)$ - оценка тематической схожести по Хэммингу; d_i – эталон; d_j – документ, сходство которого с эталоном оценивается; *c* – коллекция текстов, к которой относятся документы d_i, d_j, $\hat{W}(d_i)$; $\hat{W}(d_j)$ – множества наиболее значимых ЛД документов d_i и d_i, соответственно.

Множество $\hat{W}(d)$ формируется на основе рассчитанных по формуле (5) весов ЛД. При этом в $\hat{W}(d)$ рассматриваются только такие ЛД, вес которых превышает некоторый порог минимальной значимости ЛД, являющийся параметром алгоритма v_{min}. Этот порог выбирается на основе эмпирических соображений таким образом, чтобы при расчете сходства не потерять информацию, связанную с наиболее значимыми ЛД. Такой подход позволяет существенно сократить количество сопоставляемых ЛД, что способствует повышению скорости работы алгоритмов на практике.

Отметим, что если известно, что документы d_i, d_j относятся к тематическим классам σ_i, σ_j (возможно, совпадающим), то для более точной оценки значимости ЛД вместо формулы (5) может использоваться формула (6).

В качестве альтернативы вышеописанному методу оценки тематической связанности документов предложен метод, учитывающий семантические значения синтаксем. Учет значений синтаксем получил применение в задачах информационного поиска [8, 9]. В частности, применение такого подхода позволяет реализовать вопросно-ответный реляционно-ситуационный поиск. В своей работе мы опирались на реляционно-ситуационную модель представления текста, предложенную Г.С. Осиповым [12] как развитие принципов лингвистической теории коммуникативной грамматики Г.А. Золотовой в области информатики.

В отличие от изложенного выше метода, построенного вокруг значимости ЛД, в методе оценки тематической связанности документов с учетом семантики вместо ЛД рассматриваются синтаксемы с установленными для них семантическими значениями. В качестве синтаксем рассматривались отдельные слова (имена существительные) и именные группы в предложениях текстов документов (аналогичные ЛД) с приписанными им семантическими значениями. При таком подходе документу соответствует вектор $S(d,c) = (v(s_1,d,c), v(s_2,d,c), ..., v(s_n,d,c))$, содержащий веса $v(s_i,d,c)$ синтаксем s_i со значениями, выделенными в тексте этого документа. Синтаксем s_i представляет собой пару $s_i =< w_k, r_m >$ – лексический дескриптор и семантическое значение, соответствующее употреблению этого ЛД в тексте документа. Расчет значимости синтаксем $v(s_i,d,c)$ производится аналогично формуле (5). Таким образом, ЛД с разными семантическими значениями являются разными признаками и имеют различные значения в $S(d, c)$.

Авторами выполнена программная реализация предложенных методов и осуществлена их экспериментальная проверка.

2. Экспериментальная оценка качества работы методов

Целью экспериментальной оценки качества работы разработанных методов является сравнение между собой различных вариантов их реализации: комбинаций оценок тематического сходства и различных признаков, характеризующих документы. В данном разделе описана методика проведения экспериментов и полученные результаты.

Сравнивались следующие комбинации оценок тематического сходства и признаков документов:

1. Оценка тематического сходства – косинусная (формула (7)), набор признаков – множество ЛД документа.
2. Оценка тематического сходства – по Хэммингу (формула (8)), набор признаков – множество ЛД документа.
3. Оценка тематического сходства – по Хэммингу (формула (8)), набор признаков – множество ЛД документа с учетом их семантических значений.
4. Оценка тематического сходства – по Хэммингу (формула (8)), набор признаков – подмножество ЛД документа, состоящее только из существительных и именных групп.

В качестве тестовой коллекции был выбран массив документов, содержащий около 30 тысяч научных публикаций (20 тысяч научных статей из ведущих научных изданий и 10 тысяч авторефератов диссертаций, опубликованных в открытом доступе в Интернете). Язык текста документов – русский.

Проведение экспериментов включало следующие этапы обработки, типичные для информационно-аналитической системы:

- загрузка документов;
- выделение текста из электронного представления;

- лингвистический анализ, включающий морфологический, синтаксический и семантический.

В ходе морфологического анализа производилась нормализация однословных ЛД (приведение словоупотреблений к словарной форме), определение морфологических признаков и снятие омонимии. На этапе синтаксического анализа были выделены синтаксические группы – кандидаты в составные ЛД. Для выполнения этих этапов анализа использована система лингвистического анализа текстов – АОТ[17, 18]. Семантический анализ и установление значений синтаксем проведены с применением анализатора, представленного в работе [4].

В качестве составных ЛД рассматривались именные группы, соответствующие синтаксическим структурам, приведенным в Табл. 1.

После выполнения вышеописанных этапов предварительной обработки произведен расчёт оценок тематического сходства документов коллекции друг с другом, включающий следующие этапы:

1. Построение множеств признаков, характеризующих тексты документов.
2. Расчет частотных характеристик признаков в коллекции.
3. Расчет значимости признаков, характеризующих тексты документов.
4. Непосредственно расчёт оценок тематического сходства.
5. Ранжирование списков тематически схожих документов по убыванию полученных оценок.

В проведенных экспериментах для расчета значимости признаков использовалась формула (5). Данные об априорной тематической принадлежности документов к тем или иным тематическим рубрикам не использовались.

Оценка качества работы методов выполнена на основе мнений экспертов-асессоров. Для этого случайным образом был сформирован список документов-эталонов – набор документов из разных предметных областей. Для каждого эталона сформированы списки тематически похожих документов с применением методов 2-4 (метод 1 проверялся позже). Из каждого списка отобраны ТОП-7 наиболее похожих на эталон документов и на основе их объединения сформирован единый список найденных документов (метод общего котла [1, 7, 16]). Порядок документов в общем котле – произвольный. Для каждого эталона асессорам предлагалось оценить список похожих документов, построенный по методу общего котла. Асессоры не знали, каким методом найден тот или иной документ и какова величина оценки тематического сходства, рассчитанная системой.

Табл.1. Синтаксические структуры для выделения двусловных сочетаний-кандидатов в значимые ЛД

Синтаксическая структура	**Пример**
[Прил. + Сущ.]	Бубонная чума
[Прич. + Сущ.]	Опоясывающий лишай
[Сущ. + Сущ., Род.п.]	Рак мозга

Асессоры оценивали тематическое сходство каждого документа из списка найденных документов с эталоном по пятибалльной шкале. Примененная шкала оценки содержит следующие значения лингвистической переменной «*тематическая и содержательная схожесть документов*»:

1. «Совсем не похож» - эталон и документ из списка не имеют ничего общего (их нельзя отнести к одной предметной области, например, эталон из области психологии, а «похожий» - из области философии, и при этом они посвящены совсем разной проблематике).
2. «Скорее не похож» - эталон и документ из списка имеют мало общего (их возможно отнести к одной предметной области, но при этом они посвящены разной проблематике).
3. «Скорее похож» - эталон и документ из списка имеют общую предметную область, посвящены сходной (но не одной и той же) проблематике.
4. «Сильно похож» - эталон и документ из списка имеют много общего: из одной предметной области, посвящены одной и той же проблеме, имеют сходные методы исследования, постановки задач, и т.д.
5. «Полный дубль» - документ из списка является копией эталона или содержит копию значительных частей текста эталона, возможно, с незначительными «редакторскими» изменениями.

Всего в эксперименте участвовало 2 асессора. Каждый из них для десяти отобранных эталонов оценил 155 документов.

В результате усреднения оценок асессоров получены оценки сходства с эталоном для каждого документа в шкале 1-5. Таким образом,

получены таблицы, содержащие экспертные оценки тематического сходства эталонов и найденных тремя методами документов.

Затем списки похожих документов, построенные каждым методом, были сопоставлены с таблицами экспертных оценок, и для каждого метода рассчитаны метрики качества: normalized discounted cumulative gain [14]; общая оценка полноты.

Для вычисления метрик качества использованы формулы:

$$nDCG_{Q,S} = \frac{REL_{1,Q} + \sum_{i=2}^{S} \frac{REL_{i,Q}}{\ln(i)}}{iDCG_{Q,S}},$$

$$RECALL_{Q,S,P} = \frac{\sum_{i=1}^{S} REL_{i,Q}}{\sum_{i=1}^{P} REL_{i,Q}},$$

где $nDCG_{Q,S}$ – normalized discounted cumulative gain для запроса Q при размере списка похожих документов S; $iDCG_{Q,S}$ – *ideal DCG* – наибольшее возможное значение *DCG* для запроса Q и размера списка похожих документов S; $REL_{i,Q}$ – оцененная асессорами по шкале 1-5 величина тематического сходства документа и эталона, находящегося в списке похожих документов по запросу Q на позиции i; $RECALL_{Q,S,P}$ – общая оценка полноты для запроса Q при размере списка похожих документов S и глубине пула P.

Цель использования $nDCG_{Q,S}$ – оценка качества ранжирования документов в списке похожих документов: важно не только наличие похожего документа в списке, но и величина его оценки асессором, а также положение относительно других документов в списке. Наибольшее возможное значение *DCG* (*iDCG*) достигается при строгом упорядочивании списка похожих документов в порядке убывания оценок тематического сходства документов, выставленных асессорами.

Цель использования общей оценки полноты – оценить количество найденных документов вне зависимости от их положения в списке похожих документов. При расчете $RECALL_{Q,S,P}$ использовалось S, равное 10 (при глубине пула P, равном 7). Это было сделано с целью уменьшения влияния качества ранжирования на оценку общей полноты: некоторые тематически схожие документы могут присутствовать в списке, но за пределами ТОП-7 документов, отобранных в пул. Таким образом, $nDCG_{Q,S}$ и $RECALL_{Q,S,P}$ принимают значения от 0 до 1. Большее значение метрики качества означает, что список похожих документов соответствует ожиданию пользователя системы в большей степени.

Табл. 2. Результаты эксперимента

	nDCG		Общая оценка полноты	
	Среднее	СКО	Среднее	СКО
Косинусная оценка тематического сходства (1)	0.954656	0.064402	0.799850	0.151831
Оценка тематического сходства по Хэммингу (2)	0.980957	0.038310	0.760877	0.140013
Оценка Хэмминга с использованием семантической информации (3)	0.791393	0.398514	0.365616	0.300194
Оценка Хэмминга (ЛД – только существительные и именные группы) (4)	0.968139	0.042973	0.704206	0.169376

Для каждой метрики качества рассчитано среднее значение (методом макроусреднения [1] по всем эталонам) и среднеквадратическое отклонение (СКО). Меньшее значение СКО соответствует большей стабильности соответствующей метрики на различных эталонных документах. Результаты эксперимента приведены в Табл. 2.

Метод (1) не участвовал в формировании общего котла. Его оценка была получена позже на основе таблиц мнений асессоров по результатам работы методов (2)-(4).

Из результатов эксперимента следует, что:

– наилучшим образом тематическое сходство документов определяется на основе лексики, с учетом как однословных, так и составных ЛД;

– тип меры для расчета оценок тематического сходства (7), (8) не играет существенной роли: оценка тематического сходства по Хэммингу (метод 2) в целом дает лучший результат (наивысшее среднее и наименьшее СКО), но косинусная оценка тематического сходства

(метод 1) дает результат, сопоставимый по качеству с оценкой по Хэммингу. При этом полнота выдачи несколько выше, но порядок следования документов в списке похожих документов менее оптимальный;

– игнорирование частей речи, отличных от существительных, во множестве ЛД отрицательно влияет как на полноту, так и на качество ранжирования, а также на стабильность этих показателей (величина СКО для полноты);

– предложенный способ использования семантической информации при формировании множества признаков документа не эффективен.

Последнее утверждение говорит, однако, лишь о неверно выбранной тактике использования семантики в решаемой задаче. Опираясь на полученные данные, в будущем планируется модифицировать методы (1) и (2) так, чтобы учитывать семантическую информацию следующим способом. В качестве ЛД в текстах документов наряду с отдельными словами и именными группами предлагается учитывать связанные пары *<предикатное слово, именная синтаксема>*. В качестве критерия для выделения таких пар может служить определенное семантическое значение у именной синтаксемы, установленное с помощью предикатного слова. В качестве таких значений могут рассматриваться, например, «субъект», «объект», «ликвидатив», «делибератив» и другие [19]. Это направление является предметом отдельного исследования в будущем.

Заключение

В статье предложены подходы к определению связанности научно-технических документов на основе характеристики тематической значимости. В основе разработанных методов – различные подходы к оценке тематического сходства электронных документов. Проведенное исследование и эксперименты позволили выбрать оптимальные методы поиска тематически похожих документов для реализации в поисково-аналитической системе. На основе полученных экспериментальных данных, для реализации в информационно-аналитической системе были выбраны методы (1) и (2).

Отметим, что в проведенных экспериментах определение связанности научно-технических документов осуществлялось путем сравнения каждого документа коллекции со всеми остальными. Такой подход обладает квадратичной сложностью относительно количества документов в коллекции. Время работы одного метода на коллекции в 30 тысяч документов составляет 4-6 часов (без применения параллельной обработки), что является приемлемым только для небольших коллекций. Кроме того, этот метод не применим в случае пополняемых коллекций, поскольку пересчет близости в случае поступления нового документа придется выполнять заново для предыдущих документов. Вместе с тем, практика современных информационно-аналитических систем накладывает жесткие ограничения на время обработки пользовательских запросов и дублирование хранимых данных. Вместе с тем, разработанный метод может быть адаптирован для решения задачи поиска тематически похожих документов с применением инвертированных индексов документов. Алгоритм поиска тематически похожих документов с использованием инвертированных индексов – наиболее вычислительно эффективный [2, 13]. При этом процедура построения индекса является итеративной, т.е. индекс пополняется по мере поступления новых документов в коллекцию. Разработка структур данных, а так же параллельная реализация алгоритма поиска тематически похожих документов с применением инвертированных индексов и его интеграция в поисково-аналитическую систему являются основными задачами исследований, проводимых авторами в настоящее время.

В качестве направлений сопутствующих исследований обозначим следующие:

– Получение оценок качества работы разработанных методов в составе поисково-аналитической системы с привлечением асессоров для анализа нескольких десятков эталонных документов и соответствующих им результатов. Отметим особую трудоемкость получения этих оценок, поскольку асессору необходимо внимательно изучить как эталон, так и каждый найденный системой документ, что требует значительного времени.

– Расширение понятия ЛД тесно связано и с другими вариантами использования поисково-аналитической системы - аннотированием документов, классификацией, аналитическим сопровождением пользователя для быстрого ознакомления с темой документа.

– Оценка и представление результатов решения связанных поисково-аналитических задач, непосредственно базирующихся на предложенном методе - кластеризации коллекций, выделения научных направлений, авторских коллективов и др.

Литература

1. Агеев М., Кураленок И., Некрестьянов И. Официальные метрики РОМИП - 2006 [Электронный ресурс] – Режим доступа: http://romip.ru/romip2006/appendix_a_metrics.pdf, свободный. Проверено 14.01.2013.
2. Агеев М. С., Добров Б. В. Метод эффективного расчета матрицы ближайших соседей для полнотекстовых документов // Вестник Санкт-Петербургского Университета / СПбГУ. – СПб.: 2011. – Сер. 10. Вып. 3. – с. 72-84.
3. Вейзе А. А. О ядерных текстах и их получении путем компрессии // Проблемы текстуальной лингвистики / Под. ред. проф. В.А. Бухбиндера. — Киев, 1983.
4. Завьялова О.С., Киселев А.А., Осипов Г.С., Смирнов И.В., Тихомиров И.А. Соченков И.В. Система интеллектуального поиска и анализа информации «Exactus» на РОМИП-2010 [Электронный ресурс] – Режим доступа: http://romip.ru/romip2010/04_exactus.pdf, свободный. Проверено 14.01.2013.
5. Лотман Ю.М. Структура художественного текста. — М., 1970; Общение. Текст. Высказывание. — М., 1989.
6. Мбайкоджи Э., Драль А. А., Соченков И. В. Метод автоматической классификации коротких текстовых сообщений // Информационные технологии и вычислительные системы. – 2012. – №3. – с.93 – 102.
7. Некрестьянов И., Некрестьянова М., Нозик А. К вопросу об эффективности метода “общего котла”. [Электронный ресурс] – Режим доступа: http://rcdl.ru/doc/2005/sek9_1_paper.pdf, свободный. Проверено 14.01.2013.
8. Osipov G., Smirnov I., Tikhomirov I., Zavjalova O. Application of Linguistic Knowledge to Search Precision Improvement.//Proceedings of 4th International IEEE conference on Intelligent Systems 2008. Volume 2. - P. 17-2 - 17-5.
9. Осипов Г.С., Смирнов И.В., Тихомиров И.А. Реляционно-ситуационный метод поиска и анализа текстов и его приложения //Искусственный интеллект и принятие решений. №2.2008. - С. 3-10.
10. Тихомиров И.А., Соченков И.В. Метод динамической контентной фильтрации сетевого трафика на основе анализа текстов на естественном языке // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2008. Т. 6. № 2. С. 94-100.
11. Тихомиров И. А. Exactus Expert: поисково-аналитическая система поддержки научно-технической деятельности / Тихомиров И. А., Смирнов И. В., Соченков И. В., Девяткин Д. А., Шелманов А. О., Зубарев Д. В., Швец А. В., Лешкин А. В., Суворов Р. Е. // Труды XIII национальной конференции по искусственному интеллекту с международным участием КИИ-2012, Белгород, 2012, Том 4. С. 100 - 108.
12. Osipov G. Methods for Extracting Semantic Types of Natural Language Statements from Texts. //10th IEEE International Symposium on Intelligent Control 1995, Monterey, California, USA, Aug. 1995.
13. Frakes, William B., Baeza-Yates R. Information Retrieval: Data Structures and Algorithms. - Englewood Cliffs, New Jersey: Prentice-Hall, 1992. – 504 p.
14. Kalervo Järvelin, Jaana Kekäläinen. Cumulated Gain-based Evaluation of IR Techniques. [Электронный ресурс] - Режим доступа: http://www.sis.uta.fi/infim/julkaisut/fire/KJJK-nDCG.pdf, свободный. Проверено 23.01.2013.
15. Salton G. et al. The Smart Retrieval System Experiments in Automatic Document Retrieval. – Englewood Cliffs, New Jersey: Prentice Hall Inc. – 1971.
16. Zobel J. How reliable are the Results of Large-Scale Information Retrieval Experiment? // In Proc. of the SIGIR’98. - 1998. - pp. 307-314.
17. Сокирко А. Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ) // Дисс канд.т.н. [Электронный ресурс] - Режим доступа: http://www.aot.ru/docs/sokirko/sokirko-candid-1.html, свободный. Проверено 23.01.2013.
18. Автоматическая Обработка Текста (АОТ), // [Электронный ресурс] - Режим доступа: http://www.aot.ru, свободный. Проверено 23.01.2013.
19. Золотова Г.А., Онипенко Н.К., Сидорова М.Ю. Коммуникативная грамматика русского языка. – М. 2004. –544с.

Суворов Роман Евгеньевич. Аспирант ИСА РАН, инженер-исследователь ООО «Технологии системного анализа». Окончил Рыбинский государственный авиационный технический университет им. П.А. Соловьева в 2012 году. Автор двух научных работ. Область научных интересов: тематическая классификация текстовых документов, контентная фильтрация, интеллектуальные динамические системы. E-mail: resuvorov@gmail.com

Соченков Илья Владимирович. Инженер-исследователь ИСА РАН, научный сотрудник ООО «Технологии системного анализа». Окончил Российский университет дружбы народов в 2009 году. Автор 25 научных работ. Область научных интересов: интеллектуальные методы поиска и анализа информации, обработка больших массивов данных, защита сетей, контентная фильтрация, компьютерная лингвистика. E-mail: ivsochenkov@gmail.com