

*А. А. Басипов, О. В. Демич*

## СЕМАНТИЧЕСКИЙ ПОИСК: ПРОБЛЕМЫ И ТЕХНОЛОГИИ

*А. А. Basipov, O. V. Demich*

### SEMANTIC SEARCH: ISSUES AND TECHNOLOGIES

Рассматриваются проблемы и общий алгоритм работы семантического поиска. Описаны отличительные особенности алгоритма семантического поиска. Выявлены существующие проблемы реализации семантических поисковых систем.

**Ключевые слова:** семантический поиск, онтология, формальный запрос, запрос на естественном языке, триплеты.

Problems and general algorithm of a semantic search are considered in the paper. Distinctive features of the semantic search algorithm are described. The existing realization problems of semantic search engines are revealed.

**Key words:** semantic search, ontology, formal query, natural language query, triplets.

#### **Введение**

Поисковые системы, осуществляющие поиск по ключевым словам, обеспечивают доступ к миллиардам индексированных Интернет-страниц для тысяч пользователей. Такие явления, как полисемия (одно слово имеет несколько значений) и синонимия слов (несколько слов с одним значением) увеличивают число нерелевантных результатов, выдаваемых поисковой системой. В связи с постоянно увеличивающимся числом сайтов растет потребность в тщательном анализе контента Интернет-документов для того, чтобы свести возможность получения нерелевантных результатов к минимуму. Технологии семантической паутины предоставляют возможности для решения этой проблемы.

Целью настоящего исследования является рассмотрение существующих технологий семантического поиска и определение специфических проблем, касающихся поиска документов в семантической паутине с использованием запросов на естественном языке.

#### **Идеи семантической паутины**

Семантическая паутина (Semantic Web) является расширением традиционного Интернета и нацелена на упрощение поиска и распределения информации. Данная технология основывается на элементах, построенных с использованием стандартных языков онтологий, таких как OWL. Обычные поисковые системы основываются на поиске ключевых терминов запроса в документе и не могут использовать его смысловое значение для получения результата, поэтому сообщество исследователей семантической паутины предложило использовать семантические поисковые технологии, среди которых OntoSearch, Semantic Portals, Semantic Wikis, мультиагент P2P, семантические системы маршрутов (запросов), вопросно-ответные системы, использующие онтологии для хранения баз знаний [1].

Документ семантической паутины SWD (Semantic Web Document) можно рассматривать как набор данных, контентом которого является либо онтология, либо обычный документ, размеченный определенными тегами, взятыми из онтологии предметной области. Такие Интернет-документы могут быть распределены по множеству различных категорий, относящихся к типам онтологий, используемых для разметки документа. Примерами таких категорий являются тяжеловесная или легковесная онтологии.

Рассмотрим существующие технологии семантической паутины в контексте следующих проблем: автоматическое создание формального запроса (онтологии запроса) и получение коллекций документов, структура которых не известна заранее (распределенные и семантически разнородные данные).

Стандартный Интернет-поиск по ключевым словам базируется на поисковых технологиях, основой которых является обнаружение строкового (лексического) соответствия запрашиваемых терминов терминам, содержащимся в Интернет-документах. Обычно Интернет-поиск по ключу применяется для поиска неструктурированных Интернет-документов (текст без семантической разметки).

Наиболее популярным видом Интернет-поиска является булев поиск, основанный на обнаружении комбинаций ключевых слов, разделенных операторами AND, OR, NOT. Кроме того, существуют:

- нечеткий поиск (обработка неправильного написания и множественных чисел ключевых слов);
- поиск с использованием джокеров (Wildcard-символов) и поиск с расстоянием (синтаксический анализ документов или запрашиваемых слов);
- поиск по контексту (анализирует контент Интернет-страниц и возвращает семантический элемент страницы);
- поиск, основанный на местоположении ключевых слов (ключевые слова в заголовочных тегах Интернет-страницы более важны, чем ее контент);
- предметно-ориентированный поиск (для сужения поиска и получения более релевантных результатов используются иерархии и каталоги, категории контента);
- поиск в тезаурусе (использует различные семантические отношения, например синонимы, для получения релевантных результатов, даже если термин не представлен в документе);
- поиск, основанный на статистике, например Google PageRank.

Технология ключевого поиска может являться основой для получения SWD-документов путем сопоставления искомых слов понятиям, которые соответствуют онтологическим элементам в SWD. Такая технология применяется в семантической поисковой системе Swoogle. Отличительной ее чертой является то, что она не использует семантику в алгоритме обнаружения. Вместо этого поиск соответствий основывается на лексических методах (строковое сопоставление строк с терминами, которые соответствуют понятиям в онтологии). Для алгоритма семантического поиска лексический и синтаксический анализ сходства терминов не является существенным, важно их смысловое сходство. Для примера: соответствие между запрашиваемым термином «book» и термином документа «reserve» может быть определено верно, если смысловым значением понятия «book» является «the reservation of a ticket» (синоним). С другой стороны, соответствие между термином запроса «book» и аналогичным термином в Интернет-документе может быть неправильно определено, если их смысловые значения различаются, например термин запроса «book» означает публикацию, а термин документа «book» означает резервирование (полисемия).

Для семантического соответствия необходимо, чтобы семантические значения запроса и документа были известны. Если запрос определен формально, семантика каждого термина может быть явно определена. Таким образом, если запрос представлен как онтология (онтология запроса), то значение каждого термина как онтологического понятия раскрывается через использование семантических отношений между этим понятием и другими онтологическими терминами. Такие отношения представляют собой не только отношения «is-a», но и «part-of», «meronym», «synonym» и т. д. С другой стороны, если запрос определен неформально, например на естественном языке, семантика каждого термина в запросе должна быть как-то раскрыта. Вопрос в том, как машина сможет понять, какое смысловое значение подразумевалось в запросе, чтобы получить документ, наиболее близкий по смыслу и, следовательно, более интересный для пользователя. Интеллектуальные поисковые системы, такие как AskJeeves (технология Teoma), пытаются решить эту проблему, анализируя термины и их взаимоотношения опытным путем, используя методы обработки естественного языка или переопределяя запрос совместно с пользователями. Альтернативная методика отображает каждый термин запроса на его внутреннее значение, используя комбинацию методов индексации векторного пространства, таких как LSI (Latent Semantic Indexing), и словарь, подобный WordNet. Кроме того, для осуществления семантического отображения документ (в дополнение к запросу) должен содержать свою семантику. В случае SWD семантика документа формально и явно определяется в онтологии. В случае неструктурированных документов необходимы передовые методики работы с онтологиями для выделения их семантик и их использования для выявления зависимых документов.

### **Общий алгоритм построения системы семантического Интернет-поиска (SWSS – Semantic Web Search System)**

В этом разделе описывается общий алгоритм работы системы семантического поиска, объединяющий несколько технологий и позволяющий создать мета-движок для фильтрации

SWD-документов, возвращаемых поисковым механизмом Swoogle. Swoogle – это поисковая система, основанная на индексировании поисковым роботом SWD-документов с RDF(S)-, DAML- или OWL-синтаксисом [2]. Swoogle предоставляет методы для семантически связанных документов до выполнения запросов. Она извлекает из них метаданные и подсчитывает зависимости между документами. Хотя Swoogle в настоящее время служит в качестве SWD-индексирующей системы, используемая поисковая технология основана на поиске лексического соответствия терминов запроса и проиндексированных названий онтологических классов и свойств. Целью использования Swoogle является доказательство того, что точность поиска для простого запроса может быть улучшена, если применяется предложенный метод семантического поиска.

Алгоритм (рис. 1) позволяет автоматически преобразовывать запросы на естественном языке в формальные запросы (онтологии запроса), применяемые при онтологическом поиске/ранжировании SWD-документов. Зависимые онтологии добавляются в метод устранения неоднозначности посредством словаря WordNet, который использует механизм автоматического отображения на смысловые значения для разрешения неоднозначности терминов запроса. В этом алгоритме может быть применен любой другой словарь или тезаурус, который представляет семантические связи между терминами запроса, такие как категоризация, эквивалентность, включение и т. д.

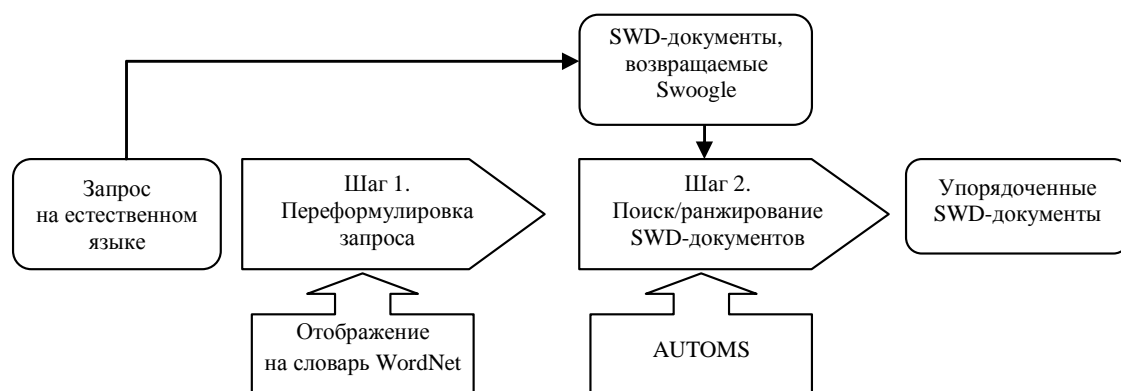


Рис. 1. Общая схема алгоритма системы семантического поиска, основанная на двухшаговом подходе (шаги отображены как блоки-стрелки)

В дополнение к автоматическому созданию онтологий запроса алгоритм использует автоматический инструмент AUTOMS для онтологического отображения. AUTOMS определяет соответствие между SWD-документом и переформулированным запросом. AUTOMS комбинирует лексические, семантические и структурные поисковые методы. Метод лексического соответствия вычисляет соответствие имен онтологических концептов, оценивает подобие понятий, используя методы для измерения синтаксической схожести. Метод структурного отображения определяет соответствие понятия онтологии с учетом терминов в его окрестности. Окрестность концепта включает те понятия, которые зависят от него. Наконец, метод семантического отображения имеет дело с сопоставлением между смысловыми значениями определений понятий. Вычисление семантического отображения может зависеть от внешней информации, найденной в словарях, тезаурусах или зависимых онтологиях.

Наконец, алгоритм позволяет проводить ранжирование полученных SWD-документов, основываясь на том, насколько хорошо они соответствуют онтологии запроса: это определяется числом соответствий между онтологией запроса и SWD-документом. Чем больше соответствий между онтологией запроса и SWD-документом, тем выше позиция SWD-документа в финальной выдаче. Фактически набор SWD-документов, который участвует в ранжирующем алгоритме, является списком документов, полученных путем передачи запроса в свободной форме в Swoogle. Таким образом, ранжирование может быть представлено как фильтрующая обработка SWD-документов, возвращаемых ключевым поиском.

Ниже представлено подробное пошаговое описание алгоритма работы системы семантического поиска, в котором изложены технические моменты для каждого отдельного шага.

**Шаг 1. Переформулирование запроса.** На этом шаге для каждого термина свободного запроса устраняется неоднозначность, оценивается предполагаемое значение, которое определяется смысловым значением WordNet. Хотя на других ступенях нашего исследования этот процесс достигается путем использования технологии LSI, в этом алгоритме мы используем модель векторного пространства (Vector Space Model – VSM) из-за характера имеющихся данных (из-за небольшого числа терминов в запросе) и необходимости уменьшить время отклика системы.

*Устранение неоднозначности запроса.* Для отображения термина запроса на соответствующее значение мы оцениваем семантическую схожесть каждого термина с множеством смысловых значений WordNet. Набор значений WordNet выбирается посредством лексического соответствия термина с контентом понятия WordNet. Алгоритм принимает во внимание окрестность  $V_t$  каждого термина запроса  $t$ .

*Векторная модель (VSM).* Термин запроса  $t$  представляется как документ («набор слов», представление). Так как термин  $t$  связан со всеми терминами в его окрестности ( $V_t$ ), документ, представляющий  $t$ , содержит все термины, встречающиеся в  $V_t$ . Кроме того, каждое значение WordNet  $S_1, S_2, \dots, S_m$ , представляющее  $m$  возможных смысловых значений термина  $t$ , представлено в виде документа.

В нашем случае мы принимаем наиболее общее представление документа в области информационного поиска, в котором весовой вектор вида  $(w_1, w_2, \dots, w_N)$ , где  $w_i, i = 1, \dots, T$ , – это *tf-idf* значение соответствующего термина  $i$ , выбранного из связанных значений WordNet или терминов запроса в окрестности  $V_t$ . Значение  $w_i$  термина  $i$  подсчитывается следующим образом:

$$w_i = tf_i \times idf_i, idf_i = \log_2 N/n_i,$$

где  $tf_i$  (term frequency, частота термина) – число вхождений термина  $i$  в отдельном документе;  $idf_i$  – инверсия числа документов, содержащих слово  $i$ ;  $N$  – общее число документов;  $n_i$  – число документов, которые содержат термин  $i$  по крайней мере один раз. Основным преимуществом использования способа *tf-idf* является то, что он позволяет определять слова, являющиеся различными для документов в корпусе. Вес слова подчеркивает его важность по отношению к другим словам, связанным с определенными документами.

Следует заметить, что в случае использования WordNet предполагаемое значение термина  $t$  определяется с помощью VSM из всех возможных значений ( $S_1, S_2, \dots, S_m$ ) соответствующего понятия WordNet. Термины выбираются из значений  $S$  WordNet с использованием следующей информации:

- слов, которые определяют смысл;
- описаний слов на естественном языке, называемых глоссарием;
- всех гиперонимов и гипонимов значения  $S$ .

Существуют случаи, когда связанная онтология может быть использована вместо словаря WordNet как внешний ресурс для устранения неоднозначности очень специализированных запросов (например, медицинских). В этом случае термины, которые будут использоваться в VSM-документах, будут выбраны из связанной онтологии с использованием следующей информации:

- названий, атрибутов и описаний онтологического понятия;
- названий, атрибутов и описаний онтологических свойств;
- названий, атрибутов и описаний понятий, зависимых по категориям или другим типам отношений.

Отображение термина запроса на документ (значение в случае WordNet или набор названий, атрибутов и описаний понятия в случае связанной онтологии) подсчитывается путем изменения расстояния между вектором запроса  $q$  и вектором каждого документа. Результатом является упорядоченный список документов. Документ с самым высоким значением косинуса подобия (cosine similarity) представляет собой предполагаемое значение термина  $t$ . Он подсчитывается следующим образом:

$$Sim(w_i, w_j) = \frac{\sum_{k=1}^T w_{ik} w_{jk}}{\sum_{k=1}^T (w_{ik})^2 \times \sum_{k=1}^T (w_{jk})^2}.$$

Шаги отображения термина запроса на значение WordNet с использованием VSM показаны ниже:

1. Выбираем термин из строки запроса. Пусть  $t$  – это название термина.
2. Получаем все значения WordNet  $S_1, S_2, \dots, S_m$ , выбранные по термину  $t$ .
3. Получаем все гиперонимы и гипонимы всех значений  $t$ .
4. Для каждого значения WordNet  $S_1, S_2, \dots, S_m$  создаем соответствующий документ, основанный на VSM.
5. Создаем документ для каждого термина запроса  $t$  с использованием всех терминов в строке запроса (т. е.  $t$  и его окрестность  $V_t$ ) с использованием VSM.
6. Находим упорядоченные связи между документом запроса термина  $t$  и документами, представляющими WordNet значения термина  $t$ , и рассматриваем связь с наибольшим косинусом подобия.

После экспериментов с другими моделями VSM (например, LSI) было обнаружено, что VSM приносит лучшие результаты, когда запрос состоит из нескольких терминов. Было протестировано несколько запросов с различным числом терминов. Был сделан вывод, что запросы с использованием трех или более терминов приносят лучший результат.

*Создание онтологии запроса.* Имея отображения терминов на значения WordNet, мы можем создать триплеты, включающие понятия и отношения между ними. В зависимости от того, что использовалось для определения предполагаемых значений терминов запроса (связанная онтология или словарь типа WordNet), будут применяться различные правила для создания онтологии запроса. Если более подробно, то для каждого термина запроса, отображенного на значение WordNet:

- создается термин, определяющий слово, которое обозначает соответствующее значение WordNet. Для примера значение термина «theater» обозначается как «dramaturgy». В результате создается термин, определенный как «dramaturgy» (рис. 2);

- если более чем одно слово определяет словарный термин, то для каждого из них создается новый термин, входящий в определение соответствующего слова. Все созданные понятия помечаются эквивалентными, поскольку все термины, обозначающие соответствующие значения WordNet, являются синонимами. Соответствующее значение термина «theater» содержит четыре синонима, а именно: «dramaturgy», «dramatic art», «dramatics», «theater» и «theatre». Как показано на рис. 2, вводятся четыре эквивалентных концепта;

- для всех гиперонимов (гипонимов) понятия WordNet создаются суперпонятия (и соответственно подпонятия) соответствующего термина на основе предыдущих правил. Результирующая таксономия для запроса «play a role in theatre» представлена на рис. 2;

- в представленном алгоритме используются два уровня гиперонимов и гипонимов. Таким образом, построенная онтология включает больше терминов, что приводит к лучшей производительности семантической системы;

- если два различных термина запроса отображаются на одно и то же значение WordNet, то оно является их общим значением и представляется в онтологии запроса как единственный концепт. Более того, если гиперонимы (или гипонимы) двух различных терминов запроса оказались одинаковыми, то созданный для него концепт соответствует одному и тому же концепту в созданной таксономии. Для примера, как показано на рис. 2, «communication» – это гипероним терминов «dramaturgy» (представляющий термин «theater») и «character» (представляющий термин «role») в таксономии WordNet. Как результат, это обобщенное понятие (суперпонятие) обоих концептов в созданной онтологии;

- другие виды семантических отношений между значениями WordNet (например, меронимы и холонимы) представлены значениями общего свойства, обозначенного как «relation». Для примера значение термина «theater» имеет только один непосредственный мероним: «dramatic composition, dramatic work (a play for performance on the stage or television or in a movie etc.)». В этом случае создается концепт, представляющий это значение посредством представленных ранее правил. Это значение связано с «theater» через «relation», как показано на рис. 2. В этом примере термин «theater» не имеет холонимов.



Таблица, приведенная ниже, представляет переранжированный список SWD-документов, возвращенных Swoogle на запрос «play theatre role». Вторая колонка таблицы представляет индекс ранжирования SWD-документов, выполненного Swoogle, третья колонка показывает индекс ранжирования, выполненного нашим алгоритмом, для которого полученная онтология запроса представлена на рис. 2. Последняя колонка показывает соотношение числа выполненных отображений к общему числу концептов в онтологии запроса.

**Точные результаты онтологии запроса «play a role in theater»  
из SWD-документов Swoogle с использованием AUTOMS**

SWD-документы (онтология)	Ранжирование Swoogle	Окончательное ранжирование	Число отображений
<a href="http://139.91.183.30:9090/RDF/VRP/Examples/DCD100.rdf">http://139.91.183.30:9090/RDF/VRP/Examples/DCD100.rdf</a>	1	1	21/52
<a href="http://athena.ics.forth.gr:9090/RDF/VRP/Examples/DCD100.rdf">http://athena.ics.forth.gr:9090/RDF/VRP/Examples/DCD100.rdf</a>	2	1	21/52
<a href="http://reliant.tekknowledge.com/DAML/Mid-level-ontology.owl">http://reliant.tekknowledge.com/DAML/Mid-level-ontology.owl</a>	3	4	4/52
<a href="http://reliant.tekknowledge.com/DAML/Mid-level-ontology.daml">http://reliant.tekknowledge.com/DAML/Mid-level-ontology.daml</a>	4	4	4/52
<a href="http://www.schemaweb.info/webservices/rest/GetRDFByID.aspx?id=241">http://www.schemaweb.info/webservices/rest/GetRDFByID.aspx?id=241</a>	5	5	3/52
<a href="http://smartweb.dfki.de/ontology/swinto0.3.1.rdfs">http://smartweb.dfki.de/ontology/swinto0.3.1.rdfs</a>	6	3	6/52
<a href="http://www.smartweb-project.org/ontology/swinto0.3.1.rdfs">http://www.smartweb-project.org/ontology/swinto0.3.1.rdfs</a>	7	3	6/52
<a href="http://www.cl.uni-heidelberg.de/kurs/ss03/ki/Uebungen/Ontologien/ontology.rdfs">http://www.cl.uni-heidelberg.de/kurs/ss03/ki/Uebungen/Ontologien/ontology.rdfs</a>	8	2	7/52

Описанный алгоритм основан на последних технологических стандартах (OWL, JENA), а также на программах автоматического отображения для нахождения соответствия онтологии запроса с SWD-документами. Общая производительность всей системы может быть оценена с точки зрения времени, точности и отзывчивости алгоритма онтологического отображения AUTOMS. Следует отметить, что время отклика системы зависит от размера SWD-документов Swoogle и может находиться в промежутке от нескольких секунд до многих минут.

### Проблемы семантического поиска

Хотя семантическая паутина способствует поиску информации в сети, существует несколько нерешенных проблем, которые следует принять во внимание. Первая из них – это огромное количество неструктурированных Интернет-документов, которые должны быть семантически размечены для использования семантическими поисковыми системами. Это непростая задача, т. к. она, среди прочего, требует развития проблемно-ориентированных онтологий.

Полностью автоматизированный процесс разметки существующих данных – еще одна нерешенная задача. С другой стороны, эффективный поиск Интернет-документов требует, вне существования онтологий, создания формальных запросов. Получается, что обычные пользователи Интернета должны изучить формальный язык для создания такого рода запросов, а это не так просто. Методы, позволяющие автоматизировать процесс преобразования запросов свободной формы (например, в форме предложения на естественном языке или как множество/список ключевых слов) к формальному виду, в настоящее время являются объектом исследования. Построение отображения онтологий предметных областей на формальные запросы также активно исследуется.

Кроме того, при разработке и реализации семантических поисковых систем возникает еще ряд проблем, которые указаны ниже.

1. Использование внешних ресурсов. SWSS должна включать дополнительные/внешние ресурсы, если для них используются запросы на естественном языке. Такое знание может быть представлено в виде общих словарей/тезаурусов или (и) в форме связанной онтологии. В этой статье мы предполагаем, что любой внешний ресурс, включающий семантическое описание терминов, может быть использован для устранения неоднозначности свободного запроса и для создания триплетов, необходимых для семантического поиска переформулированного запроса в SWD-документах. Мы также предположим, что в большинстве случаев терминология предметных областей однозначно соотносится с терминами, имеющими достаточно узкое значение.

2. Автоматизация и прозрачность. SWSS обеспечивает полностью прозрачный процесс поиска для конечного пользователя, передавая ранжированный список SWD-документов, которые находят запрос. Поиск SWD-документов осуществляется с минимумом человеческого вмешательства. Пользователям следует только проверять возвращаемую семантическим поиском информацию.

3. Производительность. Поиск SWD-документов должен выполняться быстро. Время отклика запроса в режиме реального времени в таких системах, как семантическая паутина, имеет большое значение. Таким образом, SWSS должен быть реализован как многошаговый процесс с коротким временем выполнения каждого шага для получения желаемого результата.

4. Точность/полнота. Точность SWSS-системы – тоже важный вопрос. Для запрашивания SWD-документов в реальных системах используемые технологии и реализации должны быть протестированы и оценены в контексте точности и полноты получаемого результата. В частности, автоматическое устранение неоднозначности терминов (автоматическое присваивание смысловых значений терминам) и автоматический поиск SWD-документов (отображение онтологии запроса на SWD-документы) обеспечивают более высокую точность и полноту.

### Заключение

Таким образом, нами рассмотрены проблемы семантического поиска и представлен общий алгоритм его работы. Показано, что в настоящее время существуют общие проблемы, возникающие в ходе реализации семантических поисковых систем, одной из которых является проблема автоматического преобразования запросов свободной формы в формальный вид. Представленный алгоритм работы семантической системы пока не способен полноценно обеспечивать поддержку запросов свободной формы, не требующих дополнительных навыков и знаний для выражения запросов на формальном языке.

### СПИСОК ЛИТЕРАТУРЫ

1. *Allemang D., Hendler J.* Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL // Morgan Kaufmann, 2008.
2. *Bernstein A., Kaufmann E., Fuchs N.* Talking to the semantic web – a controlled english query interface for ontologies // AIS SIGSEMIS Bulletin. – 2005. – N 2. – P. 42–47.
3. *Corcho O.* Ontology based document annotation: trends and open research problems // Int. J. Metadata, Semantics and Ontologies. – 2006. – N 1. – P. 47–57.

Статья поступила в редакцию 8.12.2011

### ИНФОРМАЦИЯ ОБ АВТОРАХ

**Басипов Андрей Алексеевич** – Астраханский государственный технический университет; аспирант кафедры «Информационные системы»; basipov@yandex.ru.

**Basipov Andrey Alekseevich** – Astrakhan State Technical University; Candidate of Technical Science; Postgraduate Student of the Department "Information Systems"; odemich@mail.ru.

**Демич Ольга Валерьевна** – Астраханский государственный технический университет; канд. техн. наук, доцент; доцент кафедры «Информационные системы»; odemich@mail.ru.

**Demich Olga Valerievna** – Astrakhan State Technical University; Candidate of Technical Science; Assistant Professor; Assistant Professor of the Department "Information Systems"; odemich@mail.ru.