Перспективы Науки и Образования. 2017. 1 (25)

Международный электронный научный журнал
ISSN 2307-2334 (Онлайн)

Адрес статьи: pnojournal.wordpress.com/archive17/17-01/
Дата публикации: 1.03.2017
№ 1 (25). С. 7-13.
УДК 001

В. Я. Цветков

# Дифференциация информационного поиска

Статья раскрывает содержание технологий информационного поиска. Статья обосновывает дифференциацию понятия релевантность результатов поиска. Статья описывает новые характеристики оценки результатов информационного поиска, включая разные оценки релевантности. Статья показывает различие между морфологической, онтологической и эпистемологической релевантностью. Показаны особенности сетевого поиска и уровни восприятия результатов поиска.

Ключевые слова: информационный поиск, сетевой поиск, паттерн, морфологическая релевантность, онтологическая релевантность, эпистемологическая релевантность



Perspectives of Science & Education. 2017. 1 (25)

International Scientific Electronic Journal
ISSN 2307-2334 (Online)

Available: psejournal.wordpress.com/archive17/17-01/
Accepted: 1 January 2017
Published: 1 March 2017
No. 1 (25). pp. 7-13.

V. Ya. Tsvetkov

# Differentiation of information retrieval

The article reveals the content of information retrieval technologies. The article substantiates the concept of differentiation of the relevance of search results. This article describes the new features evaluation of information retrieval, including different assessments of relevance. The article shows the difference between the morphological, ontological and epistemological relevance. This article describes the features of Web search and levels of perception of the search results.

Keywords: information retrieval, network search, pattern, morphological relevance, relevance ontological, epistemological relevance

## Введение

В реальной практике лицу, принимающему решение, необходима информация для решения поставленных задач. В сфере образования также возникает необходимость поиска информационных ресурсов (например, в Интернет). Одним из инструментов получения информации является информационный поиск (Information Retrieval, IR) [1]. В процессе информационного поиска проводят морфологическое и семантическое оценивание результатов поиска. Теория в области информационного поиска тесно связана с когнитивной лингвистикой. Когнитивная лингвистика как самостоятельная область современной лингвистической науки, выделилась из когнитивной науки [2]. Когнитивная лингвистика исследует сознание на материале языка. Она исследует когнитивные процессы и анализирует типы ментальных репрезентаций в сознании человека на основе применения к языку имеющихся в распоряжении лингвистики собственно лингвистических методов анализа с последующей когнитивной интерпретацией результатов исследования. В области информационного поиска существует свой язык, что делает возможным и удобным применение методов когнитивной лингвистики в этой области [3]. Чаще всего его применяют с использованием специализированных информационно-поисковых систем (ИПС) [4]. Информационный поиск решает также онтологические задачи [5].

## Механизмы информационного поиска

Поиск осуществляется на основе поискового запроса или паттерна [2]. Паттерн представляет собой морфологическую структуру, содержащую смысл или концепт [2]. Концептуально паттерн представляет собой информационную конструкцию [6, 7], которая может иметь множество реализаций в виде результатов поиска. Парадигма поиска может быть записана как:

Паттерн (концепт запроса) → ИПС (фонд поиска → Результаты поиска (концепт поиска)

Информационный поиск представляет собой процесс нахождения необходимой информации In, содержащей знания или данные, в некотором информационном множестве (фонд поиска) в соответствии с заданным критерием поиска за фиксированный интервал времени. Одному паттерну соответствует множество (до миллионов) результатов поиска. Ценность представляет не любая информация, а только та, концепт поиска которой наиболее близок к концепту запроса.

При информационном поиске смысл запроса заключается в некую символьную форму (паттерн) и представляет собой некую морфологическую информационную конструкцию. Если эта морфологическая конструкция сложная, например предложение содержащие слова в разных падежных отношениях, то в процессе поиска она, как правило, разбивается на простые информационные единицы с устранением отношений между ними. Все это создает информационную неопределенность [8] и неоднозначность результатов поиска [9]. Время поиска не сегда ограничено. Это продиктовано еще древними греками, в частности, знаменитым тропом Сиеста Эмпиритика «Об удалении в бесконечность» [10]. Механизм поиска может быть разный: циклический, рекурсивный, селективный, многоаспектно [11].

В ходе информационного поиска решаются две задачи: поиск содержательной информации (количество информации по Н. Винеру) [12] и исключение информационной неопределенности (количество информации по К.Э. Шеннону) [13]. При информационном поиске учитывают следующие факторы:

1. Информационную потребность в поиске необходимой информации In или Y.

2. Информационный поисковый язык - искусственный язык, используемый в информационно-поисковых системах с целью формализации информации, фактов и сведений.

3. Паттерн (запрос) – в общем случае шаблон с искомым образцом, составленный с использованием информационного языка.

4. Индексирование (indexing, индексация) – процесс составления или приписывания указателя (индекса) – служебной структуры данных, необходимой для последующего поиска. Выражение главного содержания текста какого-либо документа в терминах только языка информационно-поисковой системы. Применяется для упрощения поиска нужного документа среди множества других.

5. Лемматизация (lemmatization, нормализация) – приведение формы слова к словарному виду, то есть лемме [14].

6. Критерий оценки соответствия результата поиска поисковому запросу.

7. Оценка эффективности поисковой системы или эффективности технологий поиска.

8. Поисковый фонд - фонд в котором осуществляется поиск. Это множество, элементам которого ставятся во взаимно однозначное соответствие так называемые ключи (идентификаторы) - информационные элементы без внутренней структуры. Введение ключа означает установление информационного соответствия между атрибутами информационных объектов и неким новым информационным элементом.

Известны различные поисковые алгоритмы. Прямым называется поиск, алгоритм которого поиска основан на последовательном просмотре документов. Он осуществляется непосредственно по тексту документов [15], без предварительной обработки (без индексирования). Прямой поиск текста заключается в просмотре строки (слева направо) и последовательном сравнение каждой позиции с искомой подстрокой. Для этого сравнивают все символы. Прямой поиск является морфологическим.

Другие алгоритмы требуют «индексирования», предварительной обработки документов, при котором создается вспомогательный файл, «индекс», призванный упростить и ускорить поиск. Это алгоритмы инвертированных файлов, суффиксных деревьев, сигнатур [11].

## Оценка качества поиска

В результате поиска выявляется набор информационных сообщений (фонд поиска - Vr), элементы которого могут в разной степени удовлетворять субъекта, выполнявшего запрос поиска. В зависимости от соответствия между целью поиска In и элементом результата поиска Ip возможны разные ситуации, характеризуемые понятием релевантности (рис.1). Возможны три типа информационного соответствия [16] между результатом поиска и запросом: морфологическое соответствие, онтологическое соответствие, эпистемологическая соответствие. Три вида информационного соответствия обуславливают дифференциацию результатов информационного поиска и задают три типа релевантности. Эпистемологическую релевантность называют пертинентностью. [15, 17].

Формальной релевантностью (fr) или морфологической релевантностью называют соответствие поискового образа поисковому предписанию по морфологическим признакам. По этим признакам в первую очередь осуществляется от-

бор запросов и выдача результатов поиска в информационно-поисковых системах. Формальная релевантность, как правило, далека от того, что хочет получить исследователь. Одной из причин формальной релевантности является омонимия.

Омонимия семантической информационной единицы [18] состоит в том, что для нее существуют другая информационная единица с таким же морфологическим обозначением, но с другим смыслом. Например, лук (растение), лук (стрелковое оружие).

Онтологическая релевантность (or) - соответствие поискового образа поисковому предписанию по семантическим признакам, заданным в паттерне . Она предусматривает сравнение запроса In и результата поиска Ip на семантическом (смысловом) уровне. В частности, при документальном поиске сравнение происходит на естественном языке. Смысловая релевантность в меньшей степени ориентирована на формальные признаки, а больше основана на когнитивном анализе [19].

Информация, имеющая онтологическую релевантность, формируется на основе семантической операции, поэтому она несет в себе долю неопределенности. Критерий смыслового соответствия формируется человеком и устанавливает соответствие между выдаваемой информацией и смыслом запроса.

Это ставит перед исследователем дополнительную задачу - точно определить релевантный или не релевантный результат поиска. Такая задача решается когнитивными методами [19]. Онтологическая релевантность допускает нарушение морфологического соответствия. Она из причин такого нарушения синонимия. Синонимия семантической информационной единицы состоит в том, что для нее существуют другая морфологическая единица с таким же смыслом.

Онтологическая релевантность допускает неоднозначность информационного семантического соответствия. Эта неоднозначность обусловлена возможной полисемией. Полисемия семантической информационной единицы [18] состоит в том, что эта единица обладает рядом смысловых значений, актуализируемых в реальном семантическом окружении. Например, термин «Информация» является полисемическим. Он может описывать информацию в средствах массовой информации; информацию в компьютере; информацию, предаваемую насекомыми; генетическую информацию, информацию в человеческой памяти и т.д.

Эпистемологическая [20] релевантность (er) является наиболее полным критерием соответствия результатов поиска запросу. Она исключает полисемию. Иногда для повышения точности поиска используют антонимию. Антонимия семантической информационной единицы [18] – свойство, состоящее в том, что для нее существуют другая семантическая единица с противоположным смыслом (оппозиционная единица) [21].

Следует отметить, что результат поиска определяется не только правильно построенным запросом, но наличием информации о том, что необходимо искать. Как правило, в результате поиска выдается большой объем информации, которая не вся обладает формальной релевантностью (Vr). Эта информация анализируется и, если необходимо, в ней проводится уточняющий поиск. Этот анализ и дополнительный поиск уменьшают объем первоначально полученных данных и создают поле смысловой релевантности.

Результат информационного поиска целесообразно оценивать для того, чтобы сравнивать разные поисковые технологии и системы.

С с учетом рис.1 введем следующие обозначения. T – общее время поиска. Vr=fr+nr – объем поиска. Все релевантная информация - fr, нерелевантная информация - nr. Vf= Vr + dr - объем фонда, в котором выполнен поиск. dV – часть объема фонда не использованная при поиске. Объем or – объем онтологически релевантной информации. Объем er – объем эпистемологически релевантной информации. Совокупность этих характеристик приведена в таблице 1

Пояснения dfr; dor; der; dnr – остатки соответствующих характеристик в фонде, которые не вошли в результат поиска. Можно ввести следующие оценки результатов информационного поиска:

Полнота поиска по формальной релевантности Pf=fr/(fr+dfr) .

Полнота поиска по онтологической релевантности Po=or/(fr+dfr).

Полнота поиска по эпистемологической релевантности Pe=er/(fr+dfr).

Коэффициент релевантности поиска Крп = fr/ Vr.

Коэффициент релевантности фонда Крф = (fr+dfr)/ Vf.

Коэффициент онтологичности поиска Коп= or/ Vr.

Коэффициент пертинентности поиска Кпп= er/ Vr.

Скорость поиска за время T Vt= Vr/T.

Эффективность поиска за время Эп=Vr/Vf.

Эти показатели дают возможность оценивать динамические характеристики поиска и качественно оценивать результаты поиска. В отличие от многих оценок информационного поиска по одной релевантности, в данной схеме выделено три типа релевантности, которые дифференцировано позволяют оценить результаты поиска. Кроме того, введенные характеристики дают возможность оценивать эффективность разных ИПС и проводить сопоставительный анализ различных поисковых систем.

Эффективность информационного поиска измеряется совокупностью разных показателей, в



**Рис. 1.** Соотношение между эпистемологической, онтологической и формальной релевантностью

Таблица 1

Состав релевантной и не релевантной информации, выданной в результате поиска

| Выдача | Форм Релевантн. | Онтолог. релевантн. | Пертиентные | Нерелевантные | Всего |
|---|---|---|---|---|---|
| Выдано | fr | | | | fr |
| | | or | | | or |
| | | | er | | er |
| Не выдано | dfr | dor | der | dnr | dV |
| Всего | fr + dfr | or +dor | er+der | nr+dnr | Vf |

том числе технической и экономической эффективностью. Техническая эффективность информационно-поисковой системы или технологии определяется, как мера выполнять функции поиска. Экономическая эффективность поиска оценивается по стоимости выполнения этих функций. Стоимостные факторы могут изменяться с течением времени и регулироваться самим потребителем. Техническая эффективность определяется двумя группами факторов:

1. Объемно-временные характеристики: они включают объем фонда информационных массивов, объем выдачи, время поиска

2. Группа оценки полноты и точности поиска включает коэффициенты полноты (Р), релевантности (К) и эффективности (Э).

Анализ является необходимым компонентом информационного поиска, поскольку на его основе принимается решение о завершении или продолжении поиска. Существуют специальные задачи информационного поиска, решение которых позволяет расширять процесс поиска.

Информационный поиск может иметь многоаспектное представление и сводится к простому просмотру и анализу массива с результатами поиска. Например, информационный аудит является разновидностью информационного поиска, в ходе которого осуществляется информационный поиск соответствий и несоответствий нормативным документам. Технология такого поиска включает сравнение двух множеств. Это множество нормативной документации и информационного множества, описывающего реализацию некой технологии или совокупности практический действий, которая должна соответствовать этой нормативной документации.

Поисковые информационные ситуации можно моделировать по-разному. Например информационная ситуация возникает при сравнении информационной потребности Y и первоначальными данными X0, имеющимися в распоряжении пользователя. При отсутствии информационного соответствия между Y и X0, информационная ситуация в теоретико-множественном формализме отображается как

$$X0 \cap Y = \emptyset \quad (1)$$

Здесь Ø - пустое множество

Понятие информационного соответствия служит основанием поиска фонда с данными Хф, для которого имеет место

$$Y \subseteq Xф \quad (2)$$

После завершения поиска могут возникать конфликтные ситуации как ситуация несоответствия результата поиска информационным потребностям. При наличии ошибок в запросе результат поиска Хр может не соответствовать в полной мере Y. Это в формальном виде отразиться как

$$Y \neq Xp \quad (3)$$

Технология поиска включает поиск такого множества Xer, для которого концепт запроса CY соответствует концепту результата поиска CR.

В силу дифференциации возможно наличие трех концептов: формального CfR , онтологического COR и эпистемологического CER. концепту результата поиска CR содержит три концепта

$$CR = CfR + COR + CER + NR \quad (4)$$

NR – не релевантные результаты или «шум».

В ходе поиска и в результате анализа поиска требуется исключить CfR и NR. Анализировать COR и извлечь CER . Это приводит к тому, что современная технология информационного поиска сближается или сходится с технологией извлечения знаний.

## Моделирование при организации информационного поиска

Обобщенные модели информационного поиска имеют вид информационных конструкций [7]. Информационные конструкции описывают паттерн и концепт, а также позволяют осуществлять интерпретацию в информационном поле [22]. Сложность моделирования при информационном поиске в том, что приходится моделировать качественно разные сущности: процессы, объекты и ситуации. Процесс информационного моделирования при информационном поиске имеет двойственность. С одной стороны для оптимизации результата поиска надо моделировать паттерны как описательные модели. С другой стороны, для оптимизации процесса поиска надо моделировать процессы поиска. Это приводить к необходимости построения процессуальных моделей. Таким образом, моделирование в информационном поиске требует примене-

ния дескриптивных и прескриптивных моделей [23]. Разнообразие моделей по качественному и количественному признакам ставит задачу их систематизации. Основой построения современных моделей и информационных конструкций являются информационные единицы [24]. В области информационного поиска применяют поисковые информационные единицы [25].

## Эргодические аспекты поиска

Преимущество эргодических динамических систем в том, что при достаточном времени наблюдения такие системы можно описывать статистическими методами или наоборот. Эргодичность в информационном поиске проявляется в том, что поиск по одному большому идеальному фонду можно заменить поиском по нескольким реальным фондам, содержащим в совокупность ту же информацию, что и идеальный полный фонд. В реальной практике это приводит к многократным поискам по разным фондам. Другая причина многократного поиска информационная неопределенность при создании паттерна. Не всегда пользователь знает или может точно сформулировать цель запроса. Часто запрос осуществляется по косвенным признакам, что приводит к множественности результатов запроса и с большим уровнем шума.

Это приводит к необходимости многократных запросов, причем здесь часто возникает Марковская цепь, обусловленная тем, что по результатам текущего запроса, формируют следующий уточняющий запрос. Такая методика и технология дает полное основание говорить об уменьшении информационной неопределенности и с полным основанием применять теорию К.Э. Шеннона.

При многократном информационном поиске по разным порталам возникает разные ситуации. Если материал понимаем, то говорят о когнитивности, то есть о возможности восприятия и познания. При информационном поиске возможна ситуация, при которой (исходя из выражения (4)) имеет место

$$NR > CfR + COR + CER \quad (5)$$

Такая ситуация описывает шум или непознаваемость и качество портала. Выражение (5) задает уровень невосприимаемости. Характерными примером такой непознаваемой ситуации является то, что когда пользователь переходит по гиперссылке на сайт, вместо заявленного текста появляется реклама или меню желтой прессы, а заявленного тематического сообщения нет или его надо искать.

При информационном поиске по информационным порталам применяют следующие характеристики [19].

Nct - общее количество откликов информационной системы на запрос пользователя или объем запроса.

No - количество онтологических (когнитивных) запросов, полученных пользователями, из общих Nct откликов.

Ne - количество эпистемологических (когнитивных) баллов, полученных пользователями, из общих Nct откликов

Ко=No/Nct–коэффициент когнитивности является отношением количества онтологических запросов No общему количеству возможных запросов Nct Значения величины коэффициента когнитивности находятся в интервале 0≤Ко≤1;

Ке=Ne/Nct–коэффициент эпистемологичности является отношением количества эпистемологических запросов Ne общему количеству запросов Nct образовательной портальной системы. Значения величины коэффициента эпистемологичности находятся в интервале 0≤Ке≤1.

Таким образом, в процессе информационного поиска возникают четыре характеристики: характеристика объема или общего числа запросов Nct, характеристика онтологичности No, характеристика эпистемологичности Ne, характеристика не воспринимаемости NR. Эта ситуация исследована в работе [19] (рис.2)

На рис.2 представлены результаты эксперимента по исследованию эффективности сетевого информационного поиска и расчету когнитив энтропии. В качестве объекта рассматривался образовательный портал, на котором осуществлялся поиск информационных образовательных ресурсов. На графике представлены четыре кривые. Две из них предельные. Верхняя кривая (объем) характеризует максимальный объем получаемой информации. Вторая кривая (онтологичность) характеризует уровень задаваемый онтологическим концептом COR. Третья кривая (эпистомологичность) характеризует уровень, задаваемый эпистомологичным концептом CER

Нижняя кривая (невосприимаемость) характеризует уровень не восприятия сетевой информации или информационный шум NR.

## Заключение

Современные технологии информационного поиска являются технологиями, включающими информационное и когнитивное моделирование, а также технологии извлечения знаний. Современные технологии информационного поиска включают многоуровневую оценку релевантности, что требует дифференцированной оценки результатов поиска. Современные технологии информационного поиска требуют формирования разнообразных моделей в системе поиска. Технологии информационного поиска решают задачи уменьшения информационной неопределенности, и преодоления семантического разрыва. Однако в настоящее время количество исследований в этой области невелико. Поэтому данное направление требует дальнейшего исследования.

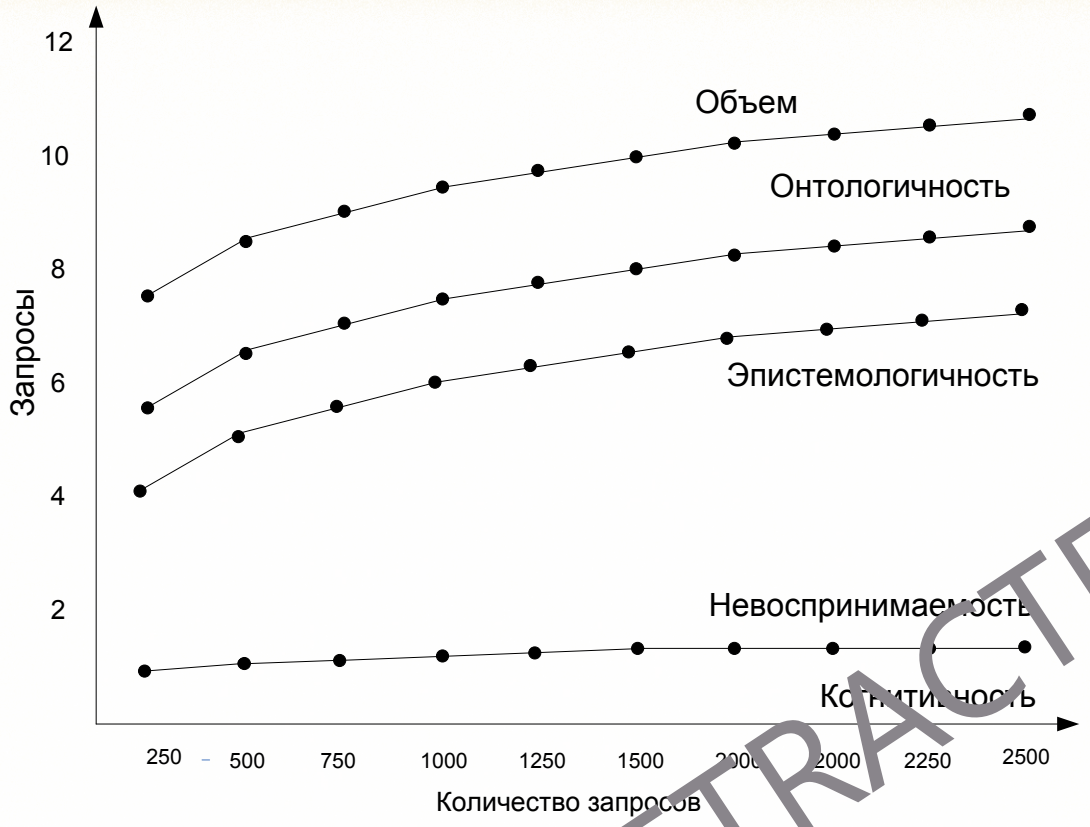


**Рис.2.** Уровни характеристик сетевого информационного поиска



## ЛИТЕРАТУРА


1. Тюрин А. Г. Математическое и программное обеспечение семантического поиска в портально-сетевых комплексах учебного назначения. Дис.к.т.н. Спец. 05.13.11. – М. МИИРЭА,2012. -121с.
2. Попова З.Д., Стернин И.А. Когнитивная лингвистика. - Москва АСТ: «Восток-Запад», 2007 -227с.
3. Tsvetkov V. Ya. Cognitive Science of Information Retrieval // European Journal of Psychological Studies, 2015, Vol.(5), Is. 1. - p.37-44. DOI: 10.13187/ejps.2015.5.37.
4. Захаров В. П. Информационно-поисковые системы – СПб, 2005.–320с.
5. Розенберг И.Н. Онтологический подход в геоинформатике // Образовательные ресурсы и технологии. – 2016. - №5 (17). – с.86-95.
6. Tsvetkov V. Ya. Information Constructions // European Journal of Technology and Design, 2014, Vol (5), № 3. - p.147-152.
7. Дешко И.П. Информационное конструирование: Монография. – М.: МАКС Пресс, 2016. – 64с. ISBN 978 -5-317-05244-7.
8. Цветков В.Я. Информационная неопределенность и определенность в науках об информации // Информационные технологии. - 2015. - №1. -с.3-7.
9. Басипов А. А., Демич О. В. Семантический поиск: проблемы и технологии //Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. – 2012. – №. 1.
10. Кибизова Э. Э. Трансцендентально-теоретическая проблематика ценности //Известия Российского государственного педагогического университета им. АИ Герцена. – 2009. – №. 109.
11. Романов В.П. Теоретические основы информатики. Информационные структуры и фактографический поиск информации.

М., Изд-во РЭА им. Г.В. Плеханова, 1996. – 190с.
12. Винер Н. Кибернетика, или управление и связь в животном и машине.1948-1961. - 2-е издание. - М.: Наука; Главная редакция изданий для зарубежных стран, 1983. - 344 с.
13. Tsvetkov V.Ya. The K.E. Shannon and L. Floridi's amount of information // Life Science Journal 2014;11 (11), pp.667-671.
14. https://en.wikipedia.org/wiki/Lemmatisation.
15. Поляков А.А., Цветков В.Я. Прикладная информатика. В 2-х частях: / Под общ.ред. А.Н. Тихонова. Том 1. - М.: МАКС Пресс. 2008. -788с.
16. Цветков В.Я. Информационное соответствие // Международный журнал прикладных и фундаментальных исследований. - 2016. - №1 (часть 3) – с.454-455.
17. Шемакин Ю.И. Теоретическая информатика. / Под общей ред. К.И. Курбакова. М.: Изд. Рос. экон. акад., 1998. - 132с.
18. Цветков В.Я. Семантика информации // Дистанционное и виртуальное обучение. 2012. № 10. С. 4-7.
19. Болбаков Р.Г. Развитие и применение когнитивно-семантических методов и алгоритмов в мультимедийных образовательных портальных системах. Дис.к.т.н. специальности 05.13.01. – М.:МИРЭА, 2013. – 131с.
20. Лекторский В.А., Кудж С.А., Никитина Е.А. Эпистемология, наука, жизненный мир человека // Российский технологический журнал 2014 - № 2 (3) - с.1-12.
21. Tsvetkov V. Ya. Opposition information analysis // European Journal of Technology and Design . – 2014. - Vol.(6), № 4, pp189-196 DOI: 10.13187/ejtd.2014.6.189.
22. Чехарин Е. Е. Интерпретация информационных конструкций // Перспективы науки и образования- 2014. - №6. – с.37-40
23. Цветков В.Я. Дескриптивные и прескриптивные информационные модели // Дистанционное и виртуальное обучение– 2015. - №7. - с.48- 54.
24. I. N. Rozenberg. Information Construction and Information Units in the Management of Transport Systems // European Journal of Technology and Design, 2016, Vol.(12), Is. 2, pp. 54-62, DOI: 10.13187/ejtd.2016.12.54 www.ejournal4.com.
25. Tajima, K., Hatano, K., Matsukura, T., Sano, R., & Tanaka, K. (1999, August).Discovery and Retrieval of Logical Information Units in Web.In WOWS (pp. 13-23.

**Информация об авторе**
**Цветков Виктор Яковлевич**
(Россия, Москва)
Профессор, доктор технических наук
Заместитель руководителя центра перспективных фундаментальных и прикладных исследований
ОАО «НИИАС»
E-mail: cvj2@mail.ru

**Information about the author**
**Tsvetkov Viktor Yakovlevich**
(Russia, Moscow)
Professor
Doctor of technical Sciences
Deputy head of the center for advanced fundamental and applied research of JSC "NIIAS"
E-mail: cvj2@mail.ru