ВИЗУАЛЬНЫЙ КОНСТРУКТОР ПОЛЬЗОВАТЕЛЬСКИХ ПОИСКОВЫХ ЗАПРОСОВ НА ОСНОВЕ КОМБИНАТОРНЫХ ТЕМАТИЧЕСКИХ ТЕЗАУРУСОВ: РЕАЛИЗАЦИЯ ИДЕИ

THE VISUAL CONSTRUCTOR OF USER'S WEB SEARCH QUERIES BASED ON COMBINATORIAL THEMATIC THESAURUSES: REALIZATION OF THE IDEA

Бочаров Алексей Владимирович,
кандидат исторических наук, доцент кафедры истории древнего мира, Средних веков и методологии истории Томского государственного университета
E-mail: bav346@rambler.ru

Alexey V. Bocharov

Предлагается концептуальная модель и техническая реализация конструктора пользовательских поисковых запросов Inspert по исторической тематике. Главная идея конструктора заключается в комбинаторно-визуальной интерпретации логических отношений между упорядоченными понятиями предметной области (дисциплинарной или отраслевой онтологии) и отправлении сконструированных сложных запросов из одного интерфейса в открытые интернет-ресурсы.

Ключевые слова: поисковый сервис, дисциплинарные онтологии и тезаурусы, методология истории, структурирование исторических знаний, визуализация, комбинаторика.

The author of this paper proposes the conceptual model and the technical implementation for the constructor of user search queries. The key concept of the constructor is to carry out the combinatorial visual interpretation of logical connections between ordered notions of a subject area (discipline- or branch-related ontology) and to send structured complex queries from a single interface to any public Internet resources. The technical implementation of the model is called *Inspert* and realized for the subject area of history.

Keywords: Search service, disciplinary ontologies and thesauri, methodology of history, of structuring of historical knowledge, visualization, combinatorics.

В научно-фантастическом романе «Осмотр на месте» (1982) Станислав Лем так описывает цивилизацию, в которой наука достигла критической стадии информационного кризиса: «…ученые все чаще приходили к убеждению, что исследуемое явление кем-то где-то наверняка подробно исследовано, неизвестно только, как найти это исследование … в университетах остались лишь компьютеры-сыщики, которые будут рыться в микропроцессорах всей планеты, чтобы узнать, ГДЕ, в каком закоулке какой машинной памяти хранятся сведения, имеющие решающее значение для проводимых исследований. <…> специалистам пришлось бы ждать от пятнадцати до шестнадцати лет, прежде чем несущаяся со скоростью света свора сигналов-ищеек успеет составить полную библиографию для задуманного исследования… началась Эпоха Экспедиций в Глубь Науки. Тех, кто планировал эти экспедиции, называли инспертами… Инсперт — эксперт на стадии заглубления (самокопания) науки…». Stanislaw Lem (pl. «Wizja lokalna», 1982). Можно считать, что во втором десятилетии XXI в. наука приближается к состоянию, описанному Лемом. Как могут и должны изменяться с учетом этого поисковые сервисы? Каким образом можно использовать историко-методологи-

ческие концепции в конструировании поисковых сервисов? Как можно организовывать интерфейсы пользователя и поисковые стратегии для планирования и проведения исторических и историографических исследований?

В качестве ответов на эти вопросы в статье предлагается концептуальная модель конструктора пользовательских поисковых запросов (constructor of user search queries). Техническая реализация модели называется Inspert, осуществлена на языке JavaScript и размещена по адресу http://www.lib.tsu.ru/inspert/. Предназначение Inspert — организация упорядоченного сбора тематической информации для научных исследований. Конструктор помогает составлять и отправлять сложные поисковые запросы в открытые для индексирования поисковиками интернет-ресурсы. Тематика первой версии конструктора — это не только историческая наука, но также «история чего угодно», или, точнее, связь любых понятий в контексте истории.

Самые популярные современные подходы к усовершенствованию тематического онлайн-поиска связаны с усложнением языка кодирования сайтов (LSI, RDF, OWL)[1] и усложнением алгоритмов их обработки (Bayesian approach for categorization)[2]. Является ли путь усложнения единственным решением? Автор предлагает исходить из того, что потенциал уже существующих поисковых технологий не исчерпан полностью и требует простых, но эвристических надстроек. Inspert основан на комбинаторно-визуальной интерпретации логических отношений между упорядоченными понятиями предметной области (дисциплинарной или отраслевой онтологии). Предлагаемая модель поиска является попыткой комплексного решения трех типов проблем: проблем пользовательского восприятия, проблем поисковых машин, проблем научных и отраслевых электронных ресурсов.

К проблемам пользовательского восприятия относятся: 1) проблема непреодолимой неосведомленности пользователей о возможностях языка поисковых запросов; 2) проблема недостаточной визуализации интерфейсов поисковых запросов; 3) проблема неудобства интерфейсов продвинутого поиска.

К проблемам поисковых машин относятся: 4) проблема избыточной и негибкой персонализации информационных фильтров в поисковых машинах (так называемый пузырь фильтров); 5) проблема недостаточной персонализации в метапоисковых машинах; 6) проблема относительной закрытости пользовательского поиска.

К проблемам научных и отраслевых электронных ресурсов относятся: 7) проблема онтологической и комбинаторной обеднённости сервисов вертикального поиска; 8) проблема ограниченности или отсутствия внутридисциплинарных тезаурусов и онтологий в навигации научных электронных библиотек; 9) проблема комбинаторной тематической неполноты сбора научной дисциплинарной или отраслевой информации.

Далее эти девять проблем онлайн-поиска будут последовательно раскрываться и сопровождаться предложениями по их решению. Реализация заявленных решений будет демонстрироваться на примере возможностей поискового сервиса Inspert. Графические интерфейсные формы для конструирования поисковых запросов названы автором клиограммами (Cliograms). Название составлено из имени Клио (древнегреческая муза истории) и слова «диаграмма».

1. ПРОБЛЕМА НЕПРЕОДОЛИМОЙ НЕОСВЕДОМЛЕННОСТИ ПОЛЬЗОВАТЕЛЕЙ О ВОЗМОЖНОСТЯХ ЯЗЫКА ПОИСКОВЫХ ЗАПРОСОВ

Общеизвестно незнание или игнорирование пользователями сложных запросов с использованием логических символов языка запросов. Практика показывает, что посредством просвещения и обучения пользователей эта проблема не решается.

Решение проблемы: a) максимизация наглядности и компактности составления сложных запросов; b) минимизация или отказ от ручного ввода специальных логических символов; c) использование систематизированных готовых шаблонов синонимических и тематических гипонимических рядов в автоматизированном генерировании запросов. Большинство пользователей не пользуются языком поисковых запросов, а если и пользуются, то не составляют сложных запросов. В частности, в исследовании K. Martzoukou (2008) подробно изучены типология поисковых стратегий и поведенческие паттерны онлайн-поиска для пользователей студентов. В этом исследовании показано, что, несмотря на общую неразвитость поисковых навыков, у молодых исследователей наблюдается рост понимания сложных стратегий поиска[3].

Проблема состоит в том, что даже умение составлять сложные запросы ограничено временем и объемом. Трудоемко вручную составлять запросы, каждый из которых состоит более чем из двадцати слов и такого же количества логических символов языка запросов. Если пользователь применяет знаки логического исключения слов из результатов поиска, то ему часто приходится тратить время для подбора именно тех слов, которые создают нерелевантный информационный шум в контексте интересующей его тематики. Гораздо удобнее, если всё

множество слов-исключений для фильтрации будет вводиться целиком и автоматически. При этом конкретное содержание множества слов-исключений обусловлено специфическими задачами научного тематического поиска.

Реализация: Важная задача при научном тематическом поиске в Интернете — нахождение всех первичных авторских текстов по теме исследования, исключив вторичные компиляции. При пользовании общедоступными поисковиками также важно исключить или существенно уменьшить навязываемые пользователю коммерческие сайты с интенсивным SEO-сервисом. Для исключения вторичных компилятивных текстов в русскоязычных поисковых запросах необходимо прежде всего исключать слова «курсовые» и «рефераты». На сайтах, распространяющих студенческие курсовые и рефераты почти нет оригинальных текстов, там содержатся скопированные и скомпилированные из чужих текстов материалы. Для исключения коммерческих сайтов необходимо исключать такие слова, как «OZON», «Amazon», «заказать», «купить», «каталог», «бесплатно», «магазин» и т. п. Контекстуальные слова-исключения связаны с конкретной тематикой поиска. Например, если поиск ведётся для исторических исследований, то необходимо, чтобы понятие «история» искалось именно в контексте терминологии истории общества, а не в контексте компьютерной терминологии «история запросов» или «кредитная история». Для более узкой тематики «культура» в рамках тематики «история» нужно исключить сайты с туристическими путеводителями, поэтому в множество слов-исключений добавляются слова «туризм», «туристический», «путеводитель», «гостиница».

Системы вопросов и ответов (Knowledge market) также не содержат научных, научно-популярных или научно-образовательных текстов, хотя вследствие популярности сервиса часто выпадают в первой десятке результатов поискового запроса. Для Рунета самой популярной справочной системой типа Knowledge market стал сервис Ответы@Mail.Ru, поэтому её название стало словом-исключением в системе Inspert. Из планируемого англоязычного поиска будут аналогично исключены названия сервисов «Ask.fm», «Spring.me», «Yahoo! Answers».

Автоматическое добавление в состав запроса синонимических рядов и гипонимических рядов также освобождает пользователя от подбора ключевых слов и логических операторов. Например, в Культурно-цивилизационной клиограмме для поиска по теме «Материальная культура» при выборе метки «Холодное оружие» в запрос автоматически добавляется тематический ряд массива слов, наиболее релевантных в результатах поисковиков тематики «Холодное оружие» по итогам разведочных запросов. Все элементы этого ряда соединяются логическим оператором «ИЛИ»: («холодное оружие» | лезвие | клинок | меч | копье | кинжал | сабля | шпага).

2. ПРОБЛЕМА НЕДОСТАТОЧНОЙ ВИЗУАЛИЗАЦИИ ИНТЕРФЕЙСОВ ПОИСКОВЫХ ЗАПРОСОВ

В поисковых сервисах, использующих визуализацию, визуализация запроса предлагается не сразу, а только при выдаче результатов исходного запроса для его расширения. Набор типичных графических схем чрезвычайно ограничен и однообразен — чаще всего это семантические гипертекстовые графы, ленты времени (timeline) или гистограммы. Более простые и интуитивно понятные диаграммы Венна не используются.

Решение проблемы: визуализация предлагается пользователю сразу, а не после появления результатов поисковой выдачи. Уже на стадии первичного формулирования запроса должны быть показаны и полностью визуализированы все поисковые логические комбинации и связи интересующей тематической области. Для достижения этой цели автор использует преимущества диаграмм Венна и Диаграммы сходства (Affinity Diagram).

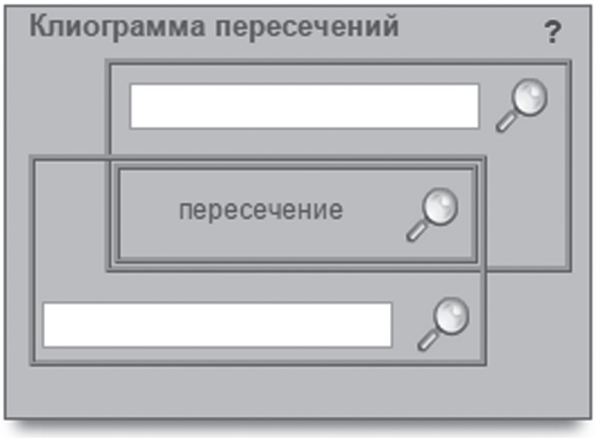

Рис. 1. Клиограмма пересечений

Реализация: клиограмма пересечений представляет собой диаграмму Венна, в которой пересечение двух фигур символизирует пересечение и разность двух множеств (complement of a sets) ключевых слов для запроса (рис. 1). Эта клиограмма в простой интерфейсной форме осуществляет возможность осуществить сразу три разных запроса. После ввода в оба поля Клиограммы пересечений можно выбрать одну из трёх кнопок поиска: либо искать пересечение введённых понятий, либо

искать одно из понятий без пересечения с другим. Под понятием подразумевается любое слово или словосочетание.

3. ПРОБЛЕМА НЕУДОБСТВА СТАНДАРТНЫХ ИНТЕРФЕЙСОВ ПРОДВИНУТОГО ПОИСКА (ADVANCED SEARCH)

Неудобство заключается: a) в относительно большом количестве форм для ввода данных с клавиатуры; b) небольшом количестве форм для выбора данных (поисковых слов и условий поиска); c) минимальном количестве кнопок запуска поиска (почти всегда только одна кнопка).

Решение проблемы: перевёртывание вышеобозначенных соотношений, т. е. создание интерфейсных форм, в которых будет: a) минимальное количество форм для ввода с клавиатуры; b) большое количество форм для выбора данных; c) максимальное количество кнопок запуска поиска, при этом каждая кнопка поиска запускает поиск с разными критериями.

Реализация: культурно-цивилизационная клиограмма (рис. 2) репрезентирует варианты кросстабуляции явлений материальной культуры и явлений духовной культуры, с азиатскими, европейскими, американскими и африканскими цивилизациями. Сферы культуры структурированы по нескольким разделам. Разделы материальной культуры: «Сферы деятельности» и «Объекты и предметы». Разделы духовной культуры: «Концепты» и «Виды искусства». В этой клиограмме 8 кнопок запуска поиска, каждая из которых генерирует отдельный поисковый запрос и 6 форм для выбора данных. Автор осознает некоторую непривычность для пользователя предложенных графических интерфейсов, однако целью при их разработке было не следование устоявшимся стандартам, а проведение эвристичного эксперимента.

В списках цивилизаций указаны только цивилизации, признание наличия которых устоялось в академической науке. Любые неустоявшиеся альтернативные названия цивилизаций, равно как и любые слова, можно вводить в поля дополнительного контекста. Если интересует история какой-то сферы материальной культуры, эту сферу можно выбрать из списка «Материальная культура», а затем, последовательно выбирая разные цивилизации, запускать поиск совместной встречаемости понятий, связанных с изучаемой сферой культуры с названиями конкретной цивилизации. У одной и той же цивилизации может быть несколько обозначений, что учитывается автоматически в формировании запроса, например, (китайская | синская | конфуцианская). Если же интересуют различные культурные аспекты одной конкретной цивилизации, то нужно выбрать её название из списка, а затем последовательно перебирать списки сфер культуры.

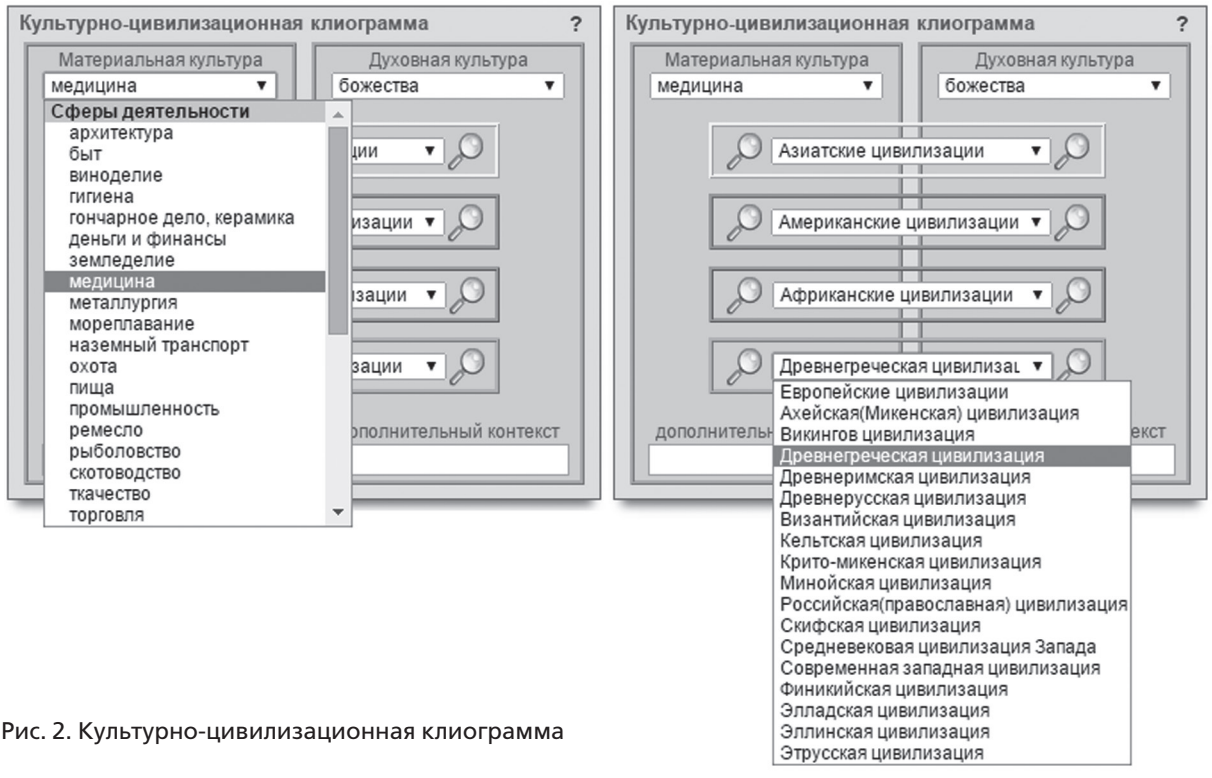

Рис. 2. Культурно-цивилизационная клиограмма

4. ПРОБЛЕМА ИЗБЫТОЧНОЙ И НЕГИБКОЙ ПЕРСОНАЛИЗАЦИИ ИНФОРМАЦИОННЫХ ФИЛЬТРОВ В ПОИСКОВЫХ МАШИНАХ («ПУЗЫРЬ ФИЛЬТРОВ»)

Понятие «Filter bubble» введено Илаем Парайзером (Eli Pariser) и описано в его книге с одноимённым названием. Парайзер критикует использование поисковиками алгоритмов выборочного угадывания того, какую информацию пользователь хотел бы увидеть, основываясь на информации об истории его персонального поиска. Поэтому в результатах поиска показывает только информацию, которая согласуется с прошлыми точками зрения данного пользователя. Вся иная информация пользователю не выводится, в результате чего он попадает в своеобразный информационный «пузырь»[4].

Нам представляется, что проблема информационной ловушки при поиске информации имеет ещё более широкий характер. Главный «пузырь фильтров» находится в наших умах. Информационная ловушка начинается, когда исследователь ограничивает поиск только узким кругом заранее известных ему параметров и признаков и отказывается от использования новых ключевых понятий, которые он не привык или его профессиональное сообщество не привыкло включать в какую-то тематику.

Решение проблемы: фильтрация информации неизбежна и необходима, но пользователь должен иметь возможность управлять ею в соответствии с тематическими смысловыми аспектами интересующей его области знаний. В современных поисковиках управлять ранжированием пользователи могут только по метаданным сайтов или файлов без учета контекстов их содержания. Исследователь не должен полностью зависеть от навязанных ему скрытых информационных фильтров, учитывающих его предыдущие, а не текущие информационные потребности. Например, поисковый сервис должен уметь искать тексты, касающиеся исторического прошлого или исторической науки, игнорируя при этом омонимическое использования слова «история».

Реализация: при составлении тезаурусов запросов в Inspert использовался дискурс-анализ. На момент запуска программа в январе 2015 г. тезаурусы содержали более 4 тыс. слов и словосочетаний по исторической и гуманитарной тематикам. Цель дискурсивного анализа — разделение тезаурусов на понятийно-терминологическую лексику и прочую, связанную с жанровыми, стилистическими и семантическими особенностями определённой предметной области. В качестве примера реализации этого решения в Inspert предлагаем рассмотреть использование антропонимов в синонимических рядах для списка по названиям идеологий и по названиям естественно-научных дисциплин. Антропонимы, т. е. наименование носителей идеологии или представителей науки, с разной вероятностью имеют прямое отношение к тематике. Упоминание «либерала» почти всегда указывает на либеральную тематику при поиске по идеологеме «либерализм». Однако упоминание «химика» или «биолога» может не иметь никакого отношения к научной тематике, а лишь указывать на профессиональную принадлежность. Поэтому дискурс-анализ на этапе формирования синонимических рядов для их автоматического включения в поисковые запросы предполагает разбор релевантности поисковой выдачи для каждого понятия из дисциплинарной таксономии и онтологии. Состав синонимических рядов, таким образом, приспосабливается к специфическим дисциплинарным контекстам, а не наоборот.

5. ПРОБЛЕМА НЕДОСТАТОЧНОЙ ПЕРСОНАЛИЗАЦИИ В МЕТАПОИСКОВЫХ МАШИНАХ

Метапоисковые машины (Metasearch engine) формируют сборную поисковую выдачу за счет смешивания результатов поиска других поисковых систем и кластеризуют найденные данные для указания на возможные направления для дальнейшего поиска. Самые известные метамашины в Рунете Quintura и Nigma, в англоязычном сегменте — Clusty. Проблема обусловлена статистическим подходом к формированию списка тегов дополнительных запросов-фильтров в метамашинах. Генерируемые метамашиной теги состоят из слов, которые чаще всего встречаются со словами, которые пользователь ввёл в своём запросе. Между тем частотный подход к фильтрациям и рекомендуемым дополнительным запросам не уменьшает информационный шум в результатах поиска, потому что часто встречающиеся слова в большинстве случаев не имеют прямого отношения к интересующей пользователя тематике. Похожие проблемы вызывают так называемые связные запросы (Related searches), предлагаемые в результатах поиска современными поисковиками. Связные запросы — это популярные запросы других пользователей, связанные с искомым запросом. Однако популярность у других пользователей никак не связана со специфическими интересами конкретного пользователя, особенно если его интересуют оригинальные тексты.

Стоит отметить, что попытки использовать инфографику для визуализации интерфейсов связ-

ных запросов уже не раз предпринимались. В 2009 г. Google добавил в функции поиск панель с расширенными возможностями визуализации поиска — колесо обозрения (wonder wheel), временну́ю шкалу (timeline). Данные сервисы предназначались для контекстуального таргетирования. Они также устанавливали связи и визуализировали их исключительно по принципу статистической частоты совместной встречаемости. Было заявлено, что эти сервисы рассчитаны на пользователя, не знающего, как уточнить запрос. Однако у широкого пользователя популярными сервисы «Google timeline» и «Google wonder wheel» не стали. Большинство их даже не замечало. Поэтому эти возможности для простых пользователей сейчас уже недоступны. С 2012 г. сервисы контекстуального таргетирования перешли к рекламным службам, таким как Google AdWords, и стали использоваться исключительно для вычисления частот совместной встречаемости слов на сайтах. Очевидно, что частотный подход к пониманию контекста непригоден для тематического поиска по научным электронным библиотекам.

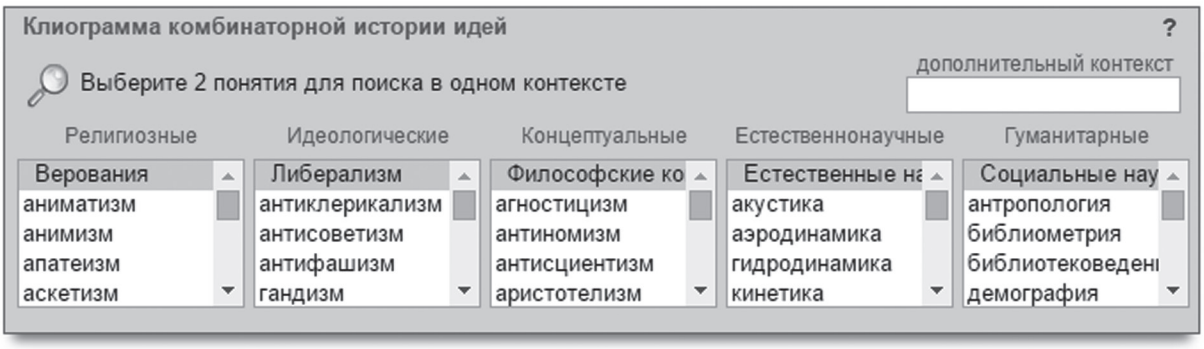

Рис. 3. Клиограмма комбинаторной истории идей

Решение проблемы: использование тематического, а не статистического подхода к формированию списка тегов для дополнительных запросов-фильтров. Тематика тегов предопределяется терминологией той области знаний, которую выбирает пользователь. Следует отметить, что Inspert предоставляет инструмент, не вмешивающийся в уже существующие способы ранжирования. Ранжирование результатов поискового запроса по-прежнему находится под управлением поисковых машин, генерирующих индексы. Inspert дает возможность углубляться в интертекст и обнаруживать смысловые, а не только гипертекстовые связи между различными понятиями.

Реализация: клиограмма комбинаторной истории идей (рис. 3) представляет собой Диаграмму сходства (Affinity Diagram), которая группирует списки названий религиозных, идеологических, концептуальных, естественно-научных и гуманитарных учений и дисциплин. В каждом списке от 120 до 200 названий и по несколько групп. Данная клиограмма позволяет составлять пары понятий из одного или из разных списков. Например, если исследователь изучает какое-либо религиозное учение, он может искать материал о связях данного учения со всеми другими религиями, идеологиями, со всеми теоретическими концепциями, гуманитарными или естественно-научными дисциплинами. Такой подход позволит определить лакуны и тенденции в огромном объеме материала.

6. ПРОБЛЕМА ОТНОСИТЕЛЬНОЙ ЗАКРЫТОСТИ ПОЛЬЗОВАТЕЛЬСКОГО ПОИСКА

Пользовательский поиск (Custom Search) содержит три опции: 1) составление создателем постоянной закрытой для пользователя подборки сайтов, внутри которой осуществляется поиск; 2) ручной ввод адресов сайтов для текущего поиска; 3) выбор сайтов из списка. Пример пользовательского поиска по русскоязычным электронным библиотекам — сайт http://tusearch.blogspot.com. На этом сайте есть фильтрация посредством пометки книг и журналов тегами «Техника», «Гуманитарные», «Математика», «Право», «Экономика», «Психология». Однако нельзя выбрать какую-то одну конкретную библиотеку и невозможно выбрать комбинацию из определённых дисциплин или выбрать только какую-то одну дисциплину без других. Такие возможности были бы актуальны для оптимизации поиска в междисциплинарных исследованиях. В результате поиск осуществляется во всех библиотеках с чрезвычайно большим количеством результатов даже после фильтрации. Когда, несмотря на фильтрацию, количество откликов в результатах поиска несколько сотен тысяч или даже несколько миллионов, становится невозможен их упорядоченный перебор. При упорядоченном поиске по множеству ре-

сурсов Всемирной паутины исследователю важно точно знать, на каких ресурсах нужные ему тексты есть, а на каких — нет. Стандартные поисковые выдачи в пользовательском поиске не сортируют результаты по сайтам и выдают их в смешанном виде. Поэтому актуальна возможность последовательного перебора ресурсов, а не только их смешивание в огромном списке, ранжированном по формальной релевантности.

Решение проблемы: преимуществом сервиса Inspert является возможность отправлять сконструированные сложные запросы из одного интерфейса в любые открытые интернет-ресурсы. Структурирование поисковой выдачи по адресам ресурсов, а не только по релевантности, позволяет лучше упорядочить процесс поиска тематической информации.

Реализация: опции направления пользовательского поиска в системе Inspert предусматривают, помимо указания любого интересующего сайта, также список полнотекстовых русскоязычных библиотек с индексируемыми текстами по исто-

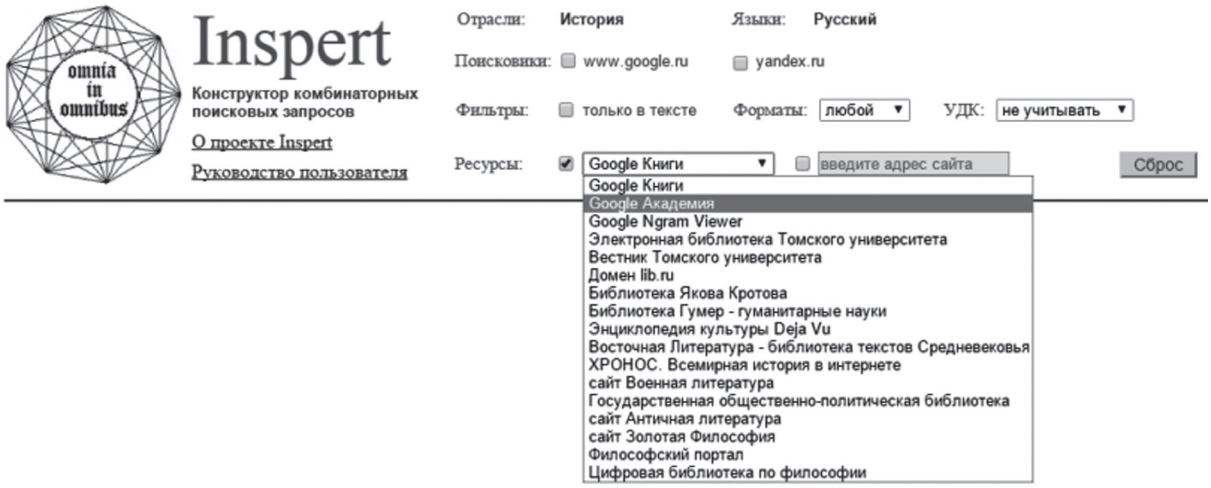

Рис. 4. Интерфейс пользовательского поиска

рической тематике (рис. 4). Для разрабатываемой мультиязычной версии сервиса, для каждого из европейских языков список библиотек будет основываться на текущих наличных ресурсах в открытом доступе. Отдельные системы вертикального поиска покрывают не все возможные научные или отраслевые ресурсы. Например, некоторые издатели не позволяли Google Scholar использовать свои полнотекстовые базы данных. Чтобы поиск был исчерпывающим, можно сконструировать в Inspert и послать в Google запрос напрямую на нужный сайт. Единственным ограничением является коммерческая политика сайта: если база данных полнотекстовых ресурсов не дает к ним открытого доступа для индексации в Google, то Inspert не поможет их увидеть.

7. ПРОБЛЕМА ОНТОЛОГИЧЕСКОЙ И КОМБИНАТОРНОЙ ОБЕДНЕННОСТИ СЕРВИСОВ ВЕРТИКАЛЬНОГО ПОИСКА

В отличие от общих поисковых машин, которые индексируют значительную часть Всемирной паутины, вертикальные поисковые системы обычно используют сфокусированный искатель, который пытается индексировать только веб-страницы, которые относятся к заранее определенной теме или набору тем. Предметно-ориентированные вертикали индексирования сосредоточиваются на одной области знания и обеспечивают чрезвычайно релевантные результаты для поисковиков из-за ограниченного корпуса доменов и более четких отношений между понятиями предметной области.

Самой известной системой вертикального поиска по научным текстам стала Google Scholar (Google-академия в русскоязычном варианте). Академия Google позволяет пользователям осуществлять поиск цифровой или физической копии статей, будь то онлайн или в библиотеках. Google Scholar является также системой индексации сводных библиографических каталогов (union catalogs). Поиск осуществляется по разным научным ресурсам, но без спецификации по дисциплинам. Дисциплинарные таксономии и онтологии в Google Scholar не используются, поэтому комбинировать различные термины поиска не получится. Такая ситуация весьма ограничивает поддержку уникальных пользовательских задач, хотя такая поддержка — главная задача вертикального поиска. Системы вертикального поиска с таксономиями и онтологиями для социаль-

ных и гуманитарных дисциплин также практически не развиваются. Исключением можно было бы считать поиск по словарям и энциклопедиям, но это поиск скорее локальный, а не вертикальный. Локальные списки гипертекстовых тегов, позволяющих искать информацию внутри сайта, не позволяют комбинировать несколько тегов и учитывать пересечения и исключения понятий.

В научной аналитике, помимо дисциплинарной онтологии терминов, важен также общенаучный аналитический тезаурус, например, абстрактные понятия, связанные с универсальными видами связей и взаимодействий. Сервисы вертикального поиска почти никогда не используют такие понятия в качестве меток для поисковых запросов, хотя именно они могут оказаться наиболее эвристичными для инновационных исследований. Исключением можно считать сервисы Technology Intelligence (TI). Сервисы TI направлены на накопление и распространение технологической информации, необходимой для стратегического планирования и принятия решений. Самым известным здесь является сервис illumin8 from ELSEVIER. В illumin8 результаты поиска группируются с помощью тезаурусов для 7 категорий взаимосвязей: Персоналии (People), Организации (Organization), Методы (Approaches), Продукты (Products), Преимущества (Benefits), Проблемы (Problems), Область исследований (Journal Keywords). Эти тезаурусы составлены главным образом для естественно-научных и технических дисциплин. Такие гуманитарные дисциплины, как история, социология, филология, психология, остались на периферии интереса разработчиков. По гуманитарным и социальным дисциплинам разработки подобных сервисов пока не актуальны для спонсоров в силу низкой капиталоёмкости гуманитарных исследований.

Решение проблемы: конструктор запросов может пониматься как внешнее расширение систем как вертикального, так и горизонтального поиска. Общенаучные понятия, связанные с универсальными видами связей и взаимодействий, не должны исчерпываться только несколькими прикладными метками типа «проблемы» и «преимущества». В интерфейсах для фильтрации результатов поиска стоит различать и учитывать гораздо больше видов взаимодействий. Их может быть не менее нескольких десятков.

Реализация: клиограмма взаимодействий позволяет искать пересечения двух разных терминов с одним типом смысловых связей. Визуально это представляет собой граф с двумя вершинами и одним ребром. В поля на вершинах графа можно вводить понятия, между которыми ищется определенный тип связи. Для запуска запроса необходимо ввести слова хотя бы в одно из двух полей (рис. 5).

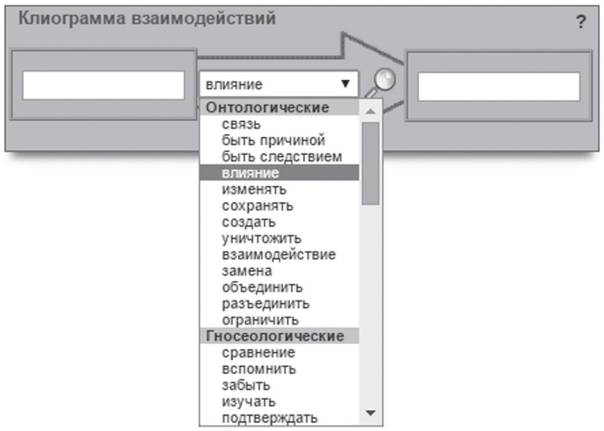

Рис. 5. Клиограмма взаимодействий

Клиограмма взаимодействий позволяет искать пересечения двух разных терминов с одним типом смысловых связей (взаимодействий). Тезаурус смысловых связей представляет третье фиксированное множество терминов. Список терминов для выбора соответствует названиям нескольких десятков видов взаимодействий.

Взаимодействия разделяются на шесть типов: Онтологические (быть причиной, быть следствием, сохранять, создать и т. п.), Гносеологические (сравнение, подтверждать, опровергать и т. п.), Аксиологические (выгодно, невыгодно, помогать, препятствовать и т. п.), Вероятностные (возможность, невозможность, неизбежность и т. п.), Темпоральные (начинать, заканчивать, ускорить, замедлить и т. п.), Мотивационные (любить, не любить, месть, зависть и т. п.). Всего более 70-ти связей. Для каждого вида связей в тезаурусе содержится синонимический ряд, уместный в контексте гуманитарных и социальных наук. Например, при выборе связи «Изменять» («Change») в поисковый запрос автоматически вводится синонимический ряд: «(изменить | менять | сменить | заменить | заменил | видоизменил | модифицировал | трансформировать | перерождаться | преображаться)».

8. ПРОБЛЕМА ОГРАНИЧЕННОСТИ ИЛИ ОТСУТСТВИЯ ВНУТРИДИСЦИПЛИНАРНЫХ ТЕЗАУРУСОВ И ОНТОЛОГИЙ В НАВИГАЦИИ НАУЧНЫХ ЭЛЕКТРОННЫХ БИБЛИОТЕК

В интерфейсах продвинутого (расширенного) поиска научных электронных библиотек обычно имеются только библиографические поля и перечни научных дисциплин, ключе-

вые слова пользователям предлагается вводить самостоятельно. Пользователь не всегда сможет скомбинировать нужные ключевые слова именно так, чтобы не упустить тексты, которые окажутся для него наиболее полезными. На текущий момент чрезвычайно мало аналогов продвижения пользовательских тематических словарей синонимов для научных дисциплин и научных электронных библиотек. Только для научно-медицинских электронных ресурсов в силу их особой актуальности наблюдаются некоторые подвижки в решении указанной проблемы. Для социальных и гуманитарных дисциплин подобных решений пока не предлагалось. Возможности детализации тематических поисковых фильтров в сервисах пользовательского поиска весьма ограничены и не удовлетворяют современным потребностям систематического полноценного поиска научной информации, ориентированного на таксономию предметной области. Кроме того, существующие системы пользовательского поиска, как правило, не предусматривают комбинирования нескольких ключевых слов и соответствующих им синонимических рядов. Уточняющие метки (Refinements labels) предназначены для выделения единичных приоритетов поиска, а не для перебора множества комбинаций.

Решение проблемы: в каждой отрасли знаний список ключевых понятий и терминов, хотя и большой, но все-таки ограниченный. Поэтому целесообразно систематизировать внутридисциплинарную терминологию и предоставлять ее в удобном виде для выбора пользователем ключевых слов с учётом гипонимии, синонимии и омонимии.

Реализация: Матрица исторического познания в табличном виде репрезентирует варианты кросстабуляции категорий исторического познания, актуальных для истории общенаучных парадигм и концепций, исторических дисциплин и субдисциплин междисциплинарной истории (рис. 6). В каждом заголовке матрицы выпадающее меню предоставляет возможность выбора из более 30 пунктов. Планирование и проведение информационной работы с помощью клиограмм предназначено для последовательной проверки связей одного интересующего пользователя ключевого понятия (кон-

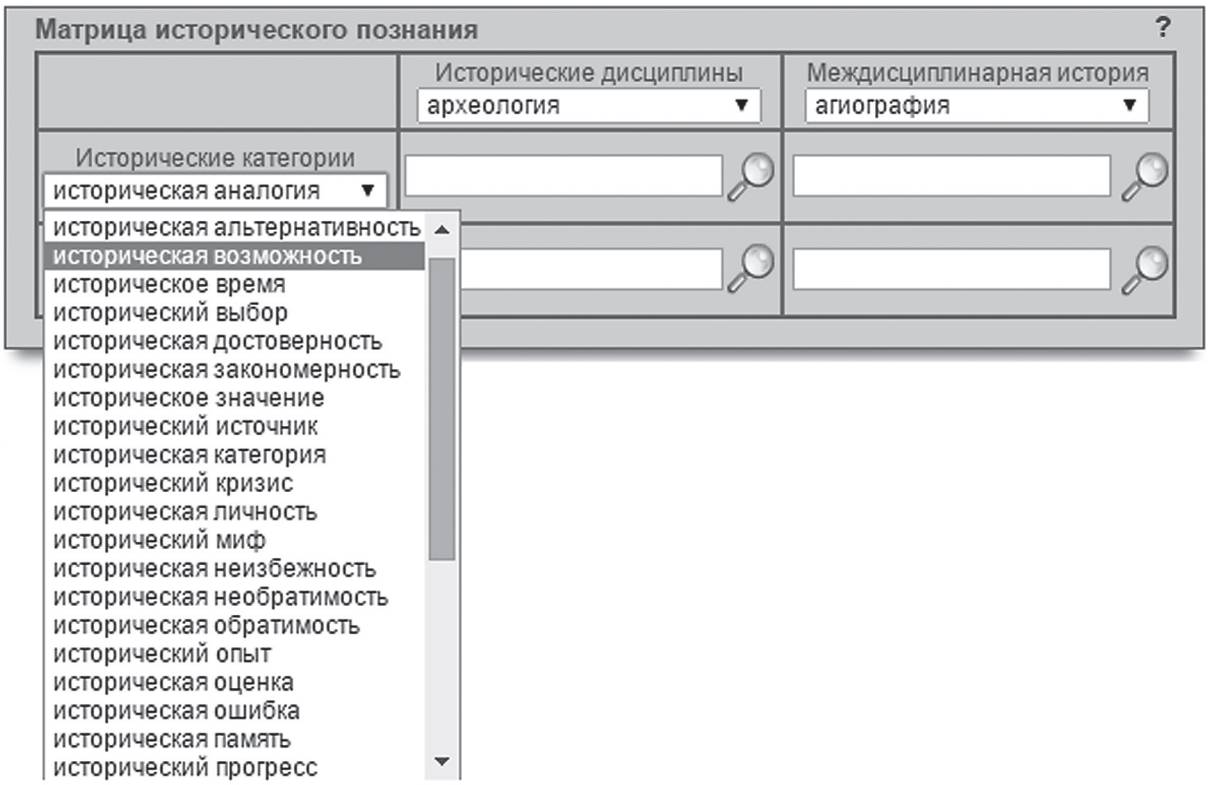

Рис. 6. Клиограмма «Матрица исторического познания»

цепта) в различных контекстах. Ключевое понятие является предметом исследования и может состоять из нескольких слов. Например, когда исследователь изучает какую-то историческую личность или событие, то он может ввести имя деятеля или название события в поля дополнительного контекста в клиограммах и установить, какие существуют связи данного имени с разными категориями исторического или общенаучного познания, а также с разными историческими субдисциплинами.

9. ПРОБЛЕМА КОМБИНАТОРНОЙ ТЕМАТИЧЕСКОЙ НЕПОЛНОТЫ СБОРА НАУЧНОЙ ДИСЦИПЛИНАРНОЙ ИЛИ ОТРАСЛЕВОЙ ИНФОРМАЦИИ

Выбор исследователя или аналитика часто обусловлен одновременно и навязанными модными трендами, и традиционными стереотипами. В результате по одним специфическим проблемам выходит избыточное количество публикаций, а по другим — очень мало или вообще ничего. Некоторые лакуны познания, логично вытекающие из предметной области и имеющихся эмпирических данных, так и остаются неосознанными или незаслуженно заброшенными.

Решение проблемы: в научно-исследовательских планах сбора информации и поиска взаимосвязей часто отсутствует предположение о симметрии в определённой системе знаний или научной дисциплине. Предположение о симметрии требует наличия недостающего элемента. Авторская концепция полноты комбинаторного научного тематического поиска основана именно на предположении о симметрии. Эвристический перебор всех возможных взаимосвязей в понятийно-терминологической онтологии какой-либо дисциплины позволяет существенно углубить и расширить любое исследование.

Реализация: описанные выше клиограммы способствуют соблюдению принципа симметрии в системе исторических знаний. Однако в представленных на данной статье клиограммах использован далеко не весь состав полного тезауруса исторического познания. При дальнейшей разработке сервиса планируется включить регионоведческую тематику — для поиска связей одного географического названия со всеми остальными терминами и названиями, просопографическую тематику — для поиска связей между именем одного исторического деятеля со всеми остальными знаменитыми именами или понятиями, методологическую тематику — для поиска связей между разными методами и методиками в контексте истории; конфликтологическую тематику — для поиска связей между разных типов исторических событий; этнографическую тематику — для поиска связей и аналогий между разными этносами; археологическую тематику — для поиска связей между специфическими археологическими терминами.

ФИЛОСОФСКИЕ ИСТОКИ МОДЕЛИ

Предложенные автором решения имеют несколько концептуальных оснований, уходящих корнями в средневековую философию и доходящих до современности. На разработку сервиса Inspert повлияли концепции нескольких европейских мыслителей. Это Раймунд Луллий, Афанасий Кирхер, Рене Декарт, Карл Манхейм.

В XIII в. Раймунд Луллий создал логический механизм в виде бумажных кругов. В своих трактатах об искусстве памяти он использовал комбинаторно-логические графические фигуры в качестве основного познавательного инструмента. Луллий предполагал, что действительность есть упорядоченное и постепенное усложнение общих понятий через их различные комбинации друг с другом, а потому разум, следя за логическим порядком понятий, может открывать действительную связь вещей[5]. Система Луллия была предназначена для работы с метафизическими универсалиями. Модификация идей Луллия в поисковом сервисе Inspert заключается в комбинаций связей между любыми абстрактными и конкретными понятыми из определённой предметной области.

Согласно замыслу Луллия, пользователь его комбинаторных вращающихся фигур должен был при переборе комбинаций понятий находить связи между ними, черпая идеи в своём собственном интеллекте и памяти. В предлагаемом здесь конструкторе поисковых запросов для нахождения связей между понятиями используются результаты интеллектуальной деятельности множества людей, чьи тексты размещены в Интернете.

В XVII в. идеи Луллия развивал Афанасий Кирхер. Он изложил комбинаторный подход к познанию в книге «Ars magna sciendi sive combinatoria» (1669). Кирхер описал проект Tabula alphabetorum artis nostrae — алфавитная таблица знаний, откуда комбинаторно можно вывести «все возможное». Пансофический проект этого великого мыслителя заключался в составлении суммы знания посредством соотнесения определенного числа базовых концептов по принципам комбинации и аналогии. Его энциклопедическая логика опирается на положение о том, что все связано друг с другом (omnia in omnibus)[6].

Первым известным призывом к полноте научного исследования можно считать высказывания Рене Декарта в его книге «Рассуждение о методе» (1606 г.) и в книге «Правила для руководства ума» (1629 г.). Именно в этих книгах можно найти многие истоки современной методологии науки. В «Правилах для руководства ума» правило VII сформулировано так: *«Чтобы придать науке полноту, надлежит все, что служит нашей цели, вместе и по отдельности обозреть в последовательном и нигде не прерывающемся движении мысли и охватить достаточной и упорядоченной нумерацией … зачастую благодаря правильно установленному порядку за короткое время и без особого труда доводится до конца многое, казавшееся на первый взгляд необъятным»*[7].

Воплотить такое «движение мысли» удобно с помощью специального поискового сервиса для

комбинирования и контаминации ключевых понятий в поисковых запросах. В структурно-функциональном анализе любой социальной или культурно-исторической системы стоит помнить, что анализ будет тем полней и результативней, чем больше связей каждого компонента системы с каждым другим компонентом будет изучено.

Более современной концепцией, повлиявшей на разработку системы Inspert, стал реляционизм. Концепт реляционизма (Relationism) предложил Карл Манхейм в книге «Идеология и утопия» (1929). По определению К. Мангейма, реляционизм — это «взаимоотнесенность всех смысловых элементов и их взаимно обосновывающая значимость внутри определенной системы»[8]. Такой системой может быть как система ценностей (идеология), так и система знаний (научная дисциплина). Мангейм отличает реляционизм от релятивизма. Принцип релятивизма означает условность и неустойчивость содержания познания. Принцип реляционизма означает взаимообусловленность и взаимовлияние всех компонентов в содержании познания. Inspert предоставляет инструмент для упорядоченного перебора понятийных связей и для раскрытия их актуального состояния в изучаемом дискурсивном поле (массиве текстов по гуманитарным тематикам).

ЗАКЛЮЧЕНИЕ

Разработка системы Inspert лежит в сфере новой области наук — инженерии знаний (knowledge engineering). Инженерия знаний изучает методы и средства извлечения, представления, структурирования и использования знаний при разработке компьютерных экспертных систем. Стоит отметить, что в социогуманитарных областях науки инженерия знаний применяется чрезвычайно редко. Локальная прикладная задача предлагаемого поискового сервиса: создание новых форм интеграции электронных ресурсов Научной библиотеки Томского государственного университета (ТГУ) в научно-исследовательские, образовательные и просветительские проекты. В дальнейшем авторский проект в рамках работы Лаборатории библиотечно-коммуникативных исследований ТГУ (НИР 8.1.39.2015) предусматривает масштабирование разработанного сервиса в мультиязычные и мультидисциплинарные формы.

Inspert позволяет структурировать как процесс, так и результаты поиска. Предлагаемый сервис — это не просто еще один интерфейс пользователя — это воплощение инновационного подхода к планированию научного исследования. Это попытка разработать своеобразную разведочную «дорожную карту» для «прокладывания» научных направлений. Одновременно — это своеобразный эксперимент по определению востребованности, перспективности форм продвижения подобных подходов к организации научно-информационного поиска в исторической науке.

Любой источник информации о современности со временем становится историческим источником. При более обобщающем взгляде любая информация — это информация о прошлом. Поэтому принципы структурирования исторического знания могут использоваться как важный пункт в проектировании систем искусственного интеллекта.

ПРИМЕЧАНИЯ

1 Almpanidis, G., Kotropoulos C., Pitas I. Focused Crawling Using Latent Semantic Indexing — An Application for Vertical Search Engines // Research and advanced technology for digital libraries: 9th european conference, ECDL 2005. Berlin; Heidelberg: Springer, 2005. P. 402–413 — (Lecture notes in computer science. — Vol. 3652).

2 Chau M., Chen H. Using Content-Based and Link-Based Analysis in Building Vertical Search Engines // Digital Libraries: International Collaboration and Cross-Fertilization 7th International Conference on Asian Digital Libraries, ICADL 2004.— 2005. P. 515–518.— (Lecture Notes in Computer Science. — Vol. 3334); Ozdikis O, Senkul P., Sinir S. Confidence-Based Incremental Classification for Objects with Limited Attributes in Vertical Search // Advanced Research in Applied Artificial Intelligence 25th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2012. Berlin; Heidelberg: Springer, 2012. P. 10–19.— (Lecture Notes in Computer Science. — Vol. 7345).

3 Martzoukou K. Students' attitudes towards web search Engines: increasing appreciation of sophisticated search strategies. Libri, 58 (43), 2008, pp.137–210.

4 Pariser E. The Filter Bubble: What the Internet Is Hiding from You, Penguin Press (New York, May 2011).

5 Bonner Anthony. The Art and Logic of Ramon Llull: A User's Guide. Brill Academic Publisher, 2007.

6 Kircher Athanasius. Ars magna sciendi sive combinatorial. Amsterdam, 1669 [Electronic resource]. Permanent URI: http://echo.mpiwg-berlin.mpg.de/MPIWG: VSKPWNWK.

7 Descartes R. Rules for the Direction of the Mind. Bobbs-Merrill Co. June 2000.

8 Mannheim, Karl. Ideology and Utopia: an Introduction to the Sociology of Knowledge / pref. author: L. Wirth, B. S. Turner; transl.: L. Wirth, E. Shils. London; New York, 1991. P. 75–76.