# СЕМИОТИЧЕСКИЙ АНАЛИЗ ЧЕЛОВЕКО-МАШИННОГО ВЗАИМОДЕЙСТВИЯ В ТЕХНОЛОГИЯХ ПОИСКА

*И.М. Зацман*

*ИПИ РАН, Москва*

Рассматриваются разработанные ранее варианты когнитивной схемы концептуального поиска документов в электронных библиотеках. Предлагается их развитие, которое заключается в реализации операций кодирования документов с использованием тезауруса на этапе подготовки автором электронных форм документов. Приведенный пример технологической схемы демонстрирует возможность организации такого кодирования при наличии он-лайнового доступа авторов к тезаурусу.

## *Введение*

Концептуальный поиск, то есть поиск на основе соответствия концептов, является в настоящее время актуальной проблемой для электронных библиотек с управляемыми информационными ресурсами и для WWW. Литературы, посвященной проблеме концептуального поиска с учетом невербальных форм представления научных знаний, почти нет. Эта проблема формулировалась применительно к текстовой информации на естественных языках, то есть рассматривались, как правило, вербальные формы представления концептов [1]. Постановка проблема поиска на основе соответствия концептов в электронных библиотеках с учетом невербальных форм представления научных знаний, включая описание когнитивных схем и задач концептуального поиска, была предложена в работе [2].

С одной стороны, поиск на основе соответствия концептов является одной из тех фундаментальных проблем информатики, разработка методов решения которых необходима для проектирования электронных библиотек новых поколений. С другой стороны, это междисциплинарная проблема, так как для ее постановки и решения, кроме теоретических основ информатики, часто необходимо привлекать и адаптировать результаты, полученные в когнитологии, семиотике и лингвистике.

Использование результатов и методов гуманитарных наук для проектирования электронных библиотек, в которых должна быть реализована возможность концептуального поиска, становится необходимым. В докладе приводится краткое описание ранее разработанных схем поиска и предлагается их развитие, основанное на результатах семиотического анализа человеко-машинного взаимодействия в процессах поиска. Предлагаемое развитие заключается в реализации операций кодирования с использованием тезауруса на этапе подготовки автором электронных форм документов. Выполнение операций кодирования на этапе подготовки документов автором возможно при наличии их электронных форм и он-лайнового доступа к тезаурусу.

## *Пример области применения*

Рассмотрим кратко пример актуальной практической проблемы, в которой задачи концептуального поиска имеют большое значение – проблема проектирования патентных электронных библиотек, предназначенных для обеспечения процесса экспертизы изобретений на новизну. Сначала кратко перечислим требования к процессу экспертизы, которые зафиксированы в Патентном законе РФ.

В статье 4 этого закона зафиксированы следующие положения: «Изобретению предоставляется правовая охрана, если оно является новым, имеет изобретательский уровень и промышленно применимо. Изобретение является новым, если оно не известно из уровня техники. Изобретение имеет изобретательский уровень, если оно для специалиста явным образом не следует из уровня техники. Уровень техники включает любые сведения, ставшие общедоступными в мире до даты приоритета изобретения. При установлении новизны изобретения в уровень техники также включаются при условии их более раннего приоритета все поданные в Российской Федерации другими лицами заявки на изобретения и полезные модели … . Экспертиза заявки на изобретение по существу включает в себя информационный поиск в отношении заявленного изобретения для определения уровня техники …».

Из приведенного перечня требований Патентного закона РФ следует ряд положений, которые являются существенными для организации процесса патентной экспертизы и проектирования патентных электронных библиотек, предназначенных для обеспечения процесса экспертизы.

Во-первых, закон требует учета любых сведений, ставших общедоступными в мире до даты приоритета заявленного изобретения, и в случае отказа в выдаче патента необходимо предоставить заявителю сведения, противопоставленные заявленному изобретению, в качестве доказательства отсутствия новизны. Очевидно, что

противопоставленные заявленному изобретению сведения должны быть в явной форме отнесены к некоторому моменту времени и этот момент должен быть раньше даты приоритета заявленного изобретения. Если в запросе на проведение поиска в электронной библиотеке используется явное указание даты приоритета заявленного изобретения, то это дает возможность не рассматривать документы электронной библиотеки, ставшие общедоступными после даты приоритета заявленного изобретения, что является стандартной задачей поиска.

Во-вторых, противопоставленные заявленному изобретению сведения должны говорить о том, что это изобретение не является новым. Важно отметить, что достаточно сложно в запросе сформулировать задачу поиска соответствия концептов заявленного изобретения содержанию уже имеющихся в электронной библиотеке описаний других изобретений. Основная причина сложности заключается в том, что в знаковых системах, как правило, действует закон асимметрии [2]. Это существенно усложняет постановку и решение задач концептуального поиска. Отметим, что частным случаем этого закона является асимметрия вербальных знаковых систем [3].

Основной целью доклада является развитие когнитивной схемы концептуального поиска документов в электронных библиотеках с использованием тезауруса, учитывающее действие закона асимметрии. Очевидно, что для электронных библиотек развитие схемы не является законченным решением задач соответствия концептов заявленного изобретения содержанию уже имеющихся в электронной библиотеке описаний других изобретений, а только одним из возможных подходов к их решению.

### *Варианты когнитивной схемы концептуального поиска*

В работе [2] рассмотрены достаточно подробно два варианта схемы. Первый вариант схемы приведен на рис. 1. На нем выделены три уровня схемы (верхний, средний и нижний) в соответствии с первыми тремя базовыми терминами концептуального поиска «знания», «информация» и «цифровые коды», которые по определению могут использоваться только на соответствующих уровнях.


***Рис. 1.*** *Когнитивная схема концептуального поиска (первый вариант)*

Второй вариант схемы приведен на рис. 2. Числами от 1 до 12 на нем обозначены 12 процессов второго варианта схемы концептуального поиска в соответствии с пятью базовыми терминами

концептуального поиска «знания», «информация», «цифровые коды», «данные» и «цифровые данные», которые по определению могут использоваться только на соответствующих уровнях.

Добавление двух терминов «данные» и «цифровые данные» увеличивает перечень процессов во втором варианте когнитивной схеме концептуального поиска с 8 до 12.


***Рис. 2.*** *Когнитивная схема концептуального поиска (второй вариант)*

В докладе различаются три основных вида концептов:

- искомые пользователем концепты (обозначим как U-концепты),
- интерпретированные пользователем концепты (I-концепты),
- авторские концепты (A-концепты).

U-концептами оперирует пользователь информационной системы при построении запросов. I-концепты являются пользовательской семантической интерпретацией найденных авторских текстовых (вербальных) и/или невербальных информационных объектов. На основании A-концептов автор создает вербальные и/или невербальные объекты.

На рис. 2 выделены пять уровней когнитивной схемы: три уровня как на рис. 1 (верхний, средний, нижний) и два новых уровня (цифровых данных и данных). Справа перечислены пять базовых терминов. Верхний уровень на рис. 2 практически совпадает с верхним уровнем на рис. 1. Отличие заключается только в том, что в явной форме не обозначены авторские и пользовательские концепты – оставлены только интерпретированный и искомый концепты.

Нижний уровень кодов и уровень цифровых данных укорочены слева, чтобы условно изобразить границу между средним уровнем информации и уровнем данных в виде точечной линии.

Приведенные варианты когнитивной схемы концептуального поиска фиксируют основные процессы, связи между ними и отношения между процессами и уровнями схемы. О методах реализации процессов эти варианты нам ничего не сообщают. Например, в запросе пользователь выражает искомый концепт, но как он это делает остается неизвестным. Однако приведенные варианты схемы помогают наглядно зафиксировать место действия закона асимметрии для знаковых систем – граница между верхним и средним уровнями.

### *Процессы когнитивной схемы*

Как отмечалось ранее, основной целью доклада является развитие схемы концептуального поиска документов в электронных библиотеках с использованием тезауруса. Развитие должно

учитывать действие закона асимметрии для следующих процессов концептуального поиска:

- представление знаний (концептов) в документах (процесс 1 на рис. 1 и 2),
- информационное представление знаний пользователей в виде запроса (формы знаков) на проведение поиска в электронной библиотеке (процесс 4 на рис. 1 и 2),
- понимание (семантическая интерпретация) пользователем найденных документов (процесс 8 на рис. 1 и 2) как отнесение значений знаков к их формам.

По определению понятия семиотического знака его значение и форма относятся, соответственно, к верхнему и среднему уровням в обоих вариантах когнитивной схемы. Далее будем рассматривать только первый вариант схемы, так как процессы 1, 4 и 8 одинаковы для обоих вариантов.

Все три процесса связаны с переходами между верхним и средним уровнями схемы. Переходы от верхнего уровня к среднему для процессов 1 и 4 условно обозначены двумя прямоугольниками «1. Создание авторских документов» и «4. Формирование информационных объектов запросов». Эти прямоугольники изображены на среднем уровне, к которому относятся результаты процессов 1 и 4. Однако основные объекты процессов 1 и 4 – концепты авторов и пользователей, информационные объекты документов и запросов - относятся к разным уровням схемы: «знания» и «информация».

Обратный переход со среднего уровня на верхний условно обозначен прямоугольником «8. Интерпретация пользователем авторских информационных объектов», т.е. семантическая интерпретация на основе восстановленных информационных объектов в его понимании. Этот прямоугольник на схеме изображен на верхнем уровне, к которому относятся результаты процесса семантической интерпретации вербальных и/или невербальных информационных объектов найденных документов. Основные объекты процесса 8 – концепты пользователя и информационные объекты найденных документов – также относятся к разным уровням схемы: «знания» и «информация».

Рассмотрим отношения объектов верхнего и среднего уровней с точки зрения семиотики, используя рис. 3. Этот рисунок создан на основе рис. 1, у которого оставлены три уровня когнитивной схемы концептуального поиска и границы между ними. По сравнению с рис. 1 удалены восемь процессов когнитивной схемы и добавлена граница между верхним и нижним уровнями, которая условно обозначена сплошной двойной линией. Уточнено также, что составной частью электронной библиотеки является система кодов элементарных концептов, знаков и форм знаков. На рис. 1 также была обозначена система кодов, но без перечисления отдельных их категорий. К первой категории по определению отнесем коды знаков, ко второй – коды их форм, а к третьей – коды их значений или элементарных концептов. На рис. 3 указаны все три категории.

Знаки на этом рисунке условно обозначены кружками на границе между верхним и средним уровнями когнитивной схемы. Именно на этой границе действует закон асимметрии. Семиотическое понятие знака как двуединой сущности, рассмотренное в контексте когнитивной схемы, привлекает внимание к границе между средним и нижним уровнями. В работе [2] по аналогии со знаком было определено новое понятие, интегрирующее форму знака и код этой формы, которое называется **формокодом**. Если использовать аналогию со знаками как элементарными единицами представления знаний в среде социальных коммуникаций, то формокоды являются элементарными единицами представления информации в цифровой среде. Было также определено еще одно новое понятие, интегрирующее элементарный концепт и его код, которое называется **семокодом**.

Важно отметить, что процессы индексирования и поиска документов относятся к цифровой среде, т.е. к нижнему уровню когнитивной схемы концептуального поиска «цифровые коды». Попытаемся изменить схему таким образом, чтобы обеспечить взаимно однозначные отношения кодов и значений знаков (концептов) в когнитивной схеме, используя понятие семокода. В тех случаях, когда выбор авторами формы представления концепта не влияет на его код в цифровой среде, исключается трансляция закона асимметрии для знаковых систем в цифровую среду.

Другими словами, если код концепта в электронной библиотеке не зависит от формы его представления на среднем уровне, то действие закона асимметрии будет ограничиваться только среднем уровнем и не будет транслироваться на нижний уровень, к которому относятся процессы индексирования и поиска документов. Отсюда следует направление развитие схемы – выбор формы представления концептов в документах электронной библиотеки и в запросах на поиск не должен влиять на коды концептов автора, принадлежащих цифровой среде.

Но реально ли хотя бы частично ограничить действие закона асимметрии среднем уровнем, если описание любого изобретения является сочетанием форм вербальных и невербальных знаков, их сочетаний, а формы по определению принадлежат к среднему уровню? Прежде чем ответить на этот вопрос, рассмотрим пример укрупненной технологической схемы подготовки электронных форм патентных заявок для формирования патентных библиотек (рис. 4). Отметим, что описания изобретений могут включать

математические и структурные химические формулы, графики, таблицы, диаграммы, чертежи и другие невербальные информационные объекты.


***Рис. 3.*** *Знаки, формокоды, семокоды и знакокоды*

### *Технологическая схема*

Технологическая схема подготовки включает четыре этапа (см. рис. 4). Первый этап имеет два входных потока – электронный и бумажный. Электронный входной поток создается авторами изобретений, которые используют программы формирования **кодированных** электронных форм патентных заявок и их подачи в соответствующее патентное ведомство с использованием Интернета в защищенном режиме передачи цифровых объектов.

В настоящее время электронные формы патентных заявок перед передачей преобразуются, как правило, в **факсимильный** вид. Поэтому технология их обработки практически полностью совпадает с технологией обработки бумажных заявок – исключается только сканирование.

Бумажный входной поток включает патентные заявки, которые создаются в бумажной форме и отправляются в патентное ведомство обычной почтой. Там они сканируются, в них разделяются вербальные и невербальные информационные объекты, после этого тексты распознаются.

На выходе первого этапа формируются три основных потока:

- изображения формул, которые передаются на второй этап технологической схемы;
- поток изображений и их подписей, который передается на третий этап;
- текстовый поток, который передается на четвертый этап технологической схемы.

Основная задача второго этапа заключается в том, чтобы расширить пространство поиска за счет формульной информации. Для патентных электронных библиотек необходимо обеспечить поиск, в первую очередь, по структурным химическим формулам. Для этого на втором этапе часть формул обрабатывается по специальным методикам. На выходе второго этапа формируется кодированный поток формульных объектов, которые передаются на четвертый этап технологической схемы.

Основная задача третьего этапа состоит в расширении пространства поиска за счет надписей и подписей к изображениям. Семантическое кодирование самих изображений, ориентированное на решение проблемы поиска, является актуальной и нерешенной задачей в технологии подготовки электронных форм патентных документов.

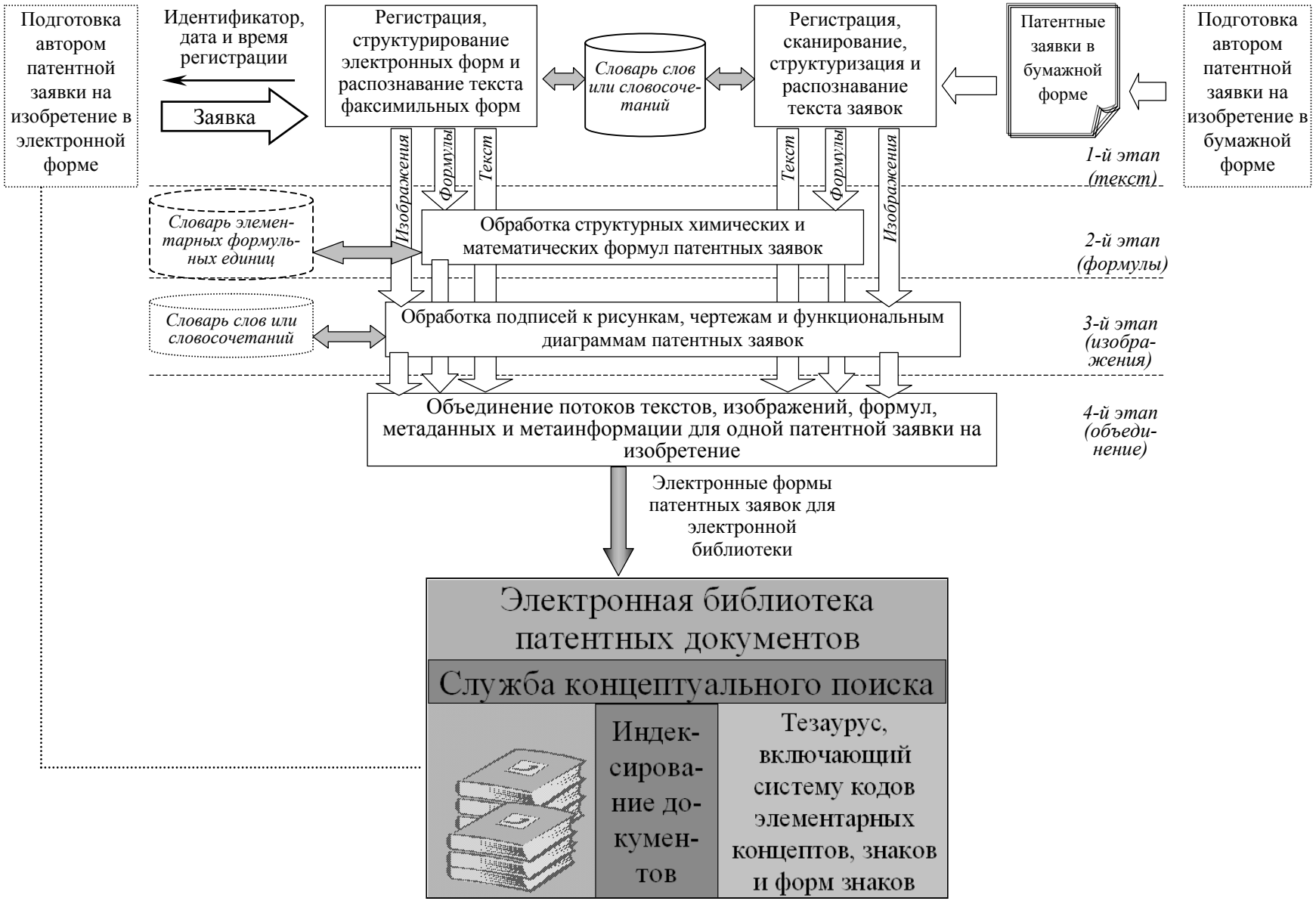


***Рис. 4.*** *Технологическая схема подготовки электронных форм патентных заявок*

Основная задача четвертого этапа заключается в объединении потоков текстов, изображений, формул, метаданных и метаинформации одного документа, а также в обработке связей между объединяемыми объектами.

После объединения компонентов электронные формы патентных заявок на изобретения становятся доступны для использования в виде целостных электронных форм документов, которые поступают в электронную патентную библиотеку.

После индексирования они становятся доступны для поиска. В докладе предполагается, что в процессах индексирования и поиска используется тезаурус [4; 5; 6], но при кодировании текста, формул и изображений в рамках технологической схемы подготовки электронных форм патентных заявок он не используется.

Индексирование электронных форм документов осуществляется после кодирования и объединения их компонентов. При выполнении операции индексирования доступны только цифровые **коды форм** вербальных и невербальных знаков, их сочетаний и/или цифровые **адреса** знаковых примитивов, например, **литер**. Но коды **значений знаков** (в авторском понимании) и кодовые выражения для авторских концептов, которые эти формы выражают, в процессе кодирования сформировать с привлечением носителей авторских концептов невозможно, так как они на этапе кодирования отсутствуют.

### *Развитие когнитивной схемы*

Предлагаемое развитие первого варианта схемы концептуального поиска (см. рис. 5) основано на выполнении операции кодирования концептов с использованием тезауруса одновременно с подготовкой автором электронных форм документов. Технологическая схема на рис. 4 иллюстрирует возможность такого использования тезауруса.

На этом рисунке возможность использования тезауруса на этапе подготовки автором электронных форм документов условно обозначена прямым углом из двух точечных линий. Предлагаемое кодирование с использованием тезауруса, выполняемое автором до отправки в патентное ведомство, предоставляет принципиально новые возможности.


***Рис. 5.*** *Когнитивная схема концептуального поиска (третий вариант)*

Во-первых, появляется возможность включить в электронные формы описаний патентных заявок одновременно все три категории кодов (см. процессы 2а-2в на рис. 5):

- коды концептов (значений знаков), являющихся уникальными идентификаторами дескрипторов тезауруса электронной библиотеки, которые формируются в результате процесса кодирования 2а,
- коды форм вербальных и невербальных знаков, которые формируются в результате процесса кодирования 2б,
- коды знаков, которые формируются в результате процесса кодирования 2в.

Во-вторых, это позволит ограничить действие закона асимметрии средним уровнем, так как дескрипторы тезауруса выбирает автор и они, в общем случае, не зависят от форм представления содержания заявляемого автором изобретения.

В-третьих, при выполнении процесса 2б появляется возможность кодирования тех форм невербальных знаков, которые не представимы в виде линейной конкатенации дискретных знаковых примитивов. Подобное представление характерно для вербальных текстов алфавитных систем письма, а для невербальных текстов линейная дискретизация в общем случае отсутствует.

В схеме на рис. 5 кодирование документов, выполняемое автором, разделено на три отдельных операции 2а, 2б и 2в в отличие от одной операции 2 на рис. 1 и 2. Кодирование запросов, выполняемое пользователем, также разделено на три операции 5а, 5б и 5в в отличие от одной операции 5 на рис. 1 и 2.

Как уже отмечалось, предлагаемое развитие схемы не является решением задач соответствия концептов заявленного изобретения содержанию уже имеющихся в электронной библиотеке описаний других изобретений, а только одним из возможных подходов к их решению. Суть предлагаемого подхода заключается в том, что символьно-кодированные и/или факсимильные электронные формы на этапе их кодирования автором дополняются кодами трех категорий: знаков, их форм и значений.

### *Заключение*

В технологической схеме подготовки электронных форм патентных заявок (рис. 4) существует два входных потока: электронный (слева) и бумажный (справа). Кодирование описаний заявок на этапе их подготовки автором возможно только при наличии их электронных форм и он-лайновом доступе к тезаурусу электронной патентной библиотеки. При практическом использовании тезауруса в процессе подготовки заявок необходимо учитывать динамику изменения тезауруса и способы фиксации динамики.

Одним из возможных способов учета динамики может быть явное указание даты и времени включения каждого нового дескриптора и установления каждой новой тезаурусной связи. Кроме указания даты и времени, необходимо с помощью специального атрибута помечать те новые дескрипторы и связи, появление которых в тезаурусе фиксирует изменение уровня техники.

Таким образом, предлагаемый способ ведения тезауруса предполагает явное указание 1) даты и времени включения дескриптора, 2) даты и времени установления тезаурусной связи, 3) значения атрибута, помечающего те дескрипторы и связи, которые фиксируют изменение уровня техники. Наличие этого атрибута позволит автору самому оценивать изобретательский уровень на этапе подготовки описания изобретения.

### *Список литературы:*


1. Schatz B.R. Information Retrieval in Digital Libraries: Bringing Search to the Net // Science Magazine. Vol. 275, No. 5298, 1997. pp. 327-334.

2. Зацман И.М. Концептуальный поиск и качество информации. М.: Наука, 2003.

3. Гак В.Г. Асимметрия // Большой энциклопедический словарь «Языкознание». М.: Большая российская энциклопедия, 1998. С. 47.

4. Шемакин Ю.И. Тезаурус в автоматизированных системах управления и обработки информации. М.: Воениздат, 1974.

5. Лукашевич Н.В. От общеполитического тезауруса к тезаурусу русского языка в контексте автоматической обработки больших массивов текстов // Труды международного семинара «Диалог 99» по компьютерной лингвистике и ее приложениям: В 2 т. Т. 2. Таруса: 1999. С. 184-190.

6. Казаков Е.Н. Формирование и ведение тезауруса в составе посредника между пользователями и сетью электронных библиотек // Труды Первой Всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции». СПб.: 1999. С. 85-88.