# Система формирования знаний в среде Интернет

# Проблемы поиска теоретической научной информации в Интернет

В настоящее время объем информационных ресурсов лавинообразно возрастает, наблюдается значительный рост количества занимаемых доменных имен. При этом также можно отметить, что возрастает интерес к сети Интернет и у коммерческих организаций, как к дополнительному источнику прибыли. Практически каждое предприятие, организация стремится иметь свое представительство в сети. В связи с этим возникают тенденции к использованию сети в коммерческих целях. Большинство российских предприятий осознают тот факт, что Интернет, в первую очередь, является дешевым средством рекламы, и воспринимают сеть, прежде всего, как рекламное печатное издание, в котором они могут разместить информацию. Только особо крупные предприятия с большим опытом работы в сети расширяют свои сайты до больших порталов, использующих средства обратной связи и общения с клиентами, ведут разъяснительную работу по поводу предоставляемых ими услуг, дают информацию, помогающую выбрать необходимую услугу или товар клиенту. Также увеличивается рост количества Интернет-магазинов и других подобных служб чисто коммерческого назначения. В связи с этими явлениями доля теоретической научной информации в общем количестве информации резко уменьшается.

не преследуют цели создать популярный ресурс. Также создатели ресурсов с научной информацией попросту слабо представляют аудиторию пользователей своих ресурсов вследствие отсутствия самых элементарных маркетинговых исследований в данной области. Также можно отметить что научно ориентированный ресурс имеет недостаточно качественное его представление и оформление. На подобных сайтах часто попросту содержатся страницы, которые являются либо документами в формате MSWord, или HTML документами, сохраненнымив том же редакторе. И как следствие имеют огромную избыточность в описании стилевых свойств и интерфейса, что уменьшает привлекательность для конечного пользователя

На данный момент, поиск научной и теоретической информации значительно затруднен вследствие некачественного ее представления в сети, а также в связи с малым объемом от общего количества информационных ресурсов и значительной ее зашумленностью.

Мониторинг и некоторые аспекты системного анализа информации в сети Интернет **выполняются с использованием** **систем информационного поиска**: разделяемые на каталоги (directories), информационно-поисковые системы (search engines) и метапоисковые системы (metasearch engines).Поисковые системы следует применять, если требуется найти информацию по специфичным вопросам или для обеспечения полноты охвата ресурсов. Многие поисковые системы используют совместно систему поиска и каталог. На данный момент основным средством поиска информации в сети Интернет являются поисковые системы. Если обобщенно рассмотреть алгоритмы работы поисковых роботов, то можно выявить следующие основные этапы их работы, которые в общем можно разделить на два параллельных процесса. **Процесс индексирования содержимого** Интернет включает:

1.Сканирование содержимого Интернет.

2.Создание индексов страниц на основании ключевых слов, заголовков и другой метаинформации.

3.Определение рейтинга индексированной страницы (**на основании цитируемости**, рейтингов сайтов, в которых встретились цитаты).

**Процесс анализа пользовательского запроса** предусматривает:

1.Проведение морфологического разбора запроса.

2.Синтаксический анализ. (Следует заметить, что для синтаксического анализа каждая система предлагает свой язык. При этом, как правило, не учитываются синтаксические конструкции того естественного языка, на котором написан запрос. В искусственном языке запросов пробелы, встречающиеся в запросе, попросту означают логическое «и», а все другие знаки пунктуации просто игнорируются. Несмотря на данные особенности, подавляющее количество пользователей поисковых систем пишет запросы на естественном языке, а многие из них даже не знают о том, что имеется встроенный язык поисковой системы, то есть большинство пользователей строят запросы с использованием только логического «и» между морфологическими основами слов.)

3.Составление запроса к базе данных, содержащей сохраненные индексы поисковых систем.

4.Определение рейтинга соответствия найденных индексов запросу пользователя. При этом учитывается лишь наиболее полное совпадение морфологических основ и близость расположения указанных основ.

5.Создание общего рейтинга на основании соответствия запросу и рейтинга популярности страницы.

6.Вывод результатов запроса в удобном для пользователя виде.

Анализ алгоритма работы поисковых систем показывает, что ни на одном из этапов не анализируется соответствие метаданных с информационным наполнением страницы. Также следует упомянуть о том, что ряд поисковых систем основывается также и на анализе форматирования страницы и содержащейся информации в заголовках страницы, разделов и других структурных единиц языка гипертекстовой разметки. Однако данные методы определения реального содержания страницы хорошо известны и могут быть легко использованы для обмана поисковых роботов, что в значительной мере используется разработчиками мелких коммерческих проектов, засоряющих результаты поиска.

Для определения цитирования информации применяются алгоритмы, которые не способны отследить перефразированные цитаты и прямое частичное зеркалирование. Большинство применяемых алгоритмов могут определить только полные зеркала сайтов. И как следствие большинство существующих поисковых систем в слабой мере агрегируют результаты поиска. Это является особо актуальной проблемой, так как реальное содержимое сети Интернет многократно дублируется. Содержимое страниц может повторяться как частично, так и полностью с незначительными изменениями.

Обобщая данные недостатки, можно отметить, что **рейтинги, получаемые при анализе страниц, слабо зависят от реального содержимого страницы** и, в основном зависят от метаданных и описания страниц, созданных самими разработчиками.

Еще один недостаток при подобном подходе к определению рейтинга заключается в том, что система оценки рейтингов зацикливается и приводит поисковую систему к статичному состоянию, что, очевидно, приводит к ее недееспособности. Если более подробно рассмотреть этот процесс, то из алгоритма определения рейтингов видно, что сайт, однажды получивший высокий рейтинг, «тяжело» теряет его и наоборот вновь появившийся сайт, не имеющий целенаправленной маркетинговой политики и не продвигаемый целенаправленно на повышение рейтинга, никогда не будет найден. Происходит данное явление вследствие того, что высокий рейтинг сайта вызывает большой приток посетителей, переходящих по первым ссылкам результатов и тем самым еще больше увеличивающих рейтинг сайтов с высоким рейтингом. А цитирование сайта еще выше поднимает рейтинг данного ресурса. Как следствие вновь разработанный ресурс, не имеющий высокого рейтинга, никогда не попадет в первые строчки поиска, несмотря на его содержание. Более того, некоторые популярные издания при оценке поисковых систем приводили огромное количество случаев, когда общий рейтинг превышал рейтинг соответствия. Вследствие чего сайт, имеющий полное соответствие запросу, но имеющий низкий рейтинг, оказывался ниже в результатах запросов, чем сайты, абсолютно не имеющие отношения к тематике запроса, но с более высоким рейтингом. Таким образом, результаты поиска становятся статичными и не способны реагировать на изменяющуюся окружающую среду Интернет.

Для увеличения рейтинга ресурса и тем самым обеспечения появления ссылки на первых страницах результатов поиска используется ряд приемов, позволяющих "обмануть" роботов, индексирующих страницы.

Разработчики стремятся расширить число запросов, которым может соответствовать их ресурс. В частности, для того, чтобы повысить рейтинг в известной поисковой системе, разработчики информационного ресурса могут **вводить в метаданные расширенную информацию, которая может вовсе не относиться к тематике их сайта и будет лишь отражением заведомо популярных тем в Интернет**, таких, как скачивание музыки, популярных программ и т.п., а также введением ключевых слов паразитов, таких, как бесплатно, скачать и др.

Еще один прием, используемый при разработке Web-ресурсов, направленна увеличение общего рейтинга страницы. Многие разработчики ресурсов могут рассылать ссылки в форумы авторитетных изданий, давать неверные ссылки читателям блогов, форумов и гостевых книг, создавать сайты–близнецы, указывающие друг на друга и приводящие цитаты с сайтов своих близнецов, проводить целенаправленный обмен ссылками между «раскручиваемыми» сайтами. Все эти приемы объединяет одна цель – увеличить цитируемость и как следствие общий рейтинг сайта в поисковой системе. С помощью подобных приемов можно увеличить цитируемость сайта и значимость его с точки зрения системы оценки рейтинга, а также частоту попадания в запросы пользователей и как следствие случайных переходов пользователей на страницы своего сайта. Подобных приемов существует достаточно много,и они не просто часто применяются в среде Интернет, а стали уже стандартом сопровождения сайтов малого и среднего бизнеса.

**Одним из решений данных проблем** является внедрение формата RSS, RDF для представления и определения метаданных, описывающих ресурсы WEB.

**Главное отличие семантического web’а от среды WWW** заключается в том, что в нем страницы содержат информацию на двух языках: обычном, понятном человеку и показываемом браузером, и специальном (описываемом онтологиями), информация на котором скрыта от людей, содержит семантическую составляющую и предназначена для различных программ, агентов и роботов.

# Формализованное описание и классификация методов мониторинга и системного анализа информации

Мониторинг информации определенной направленности является частью любого средства информационного поиска Интернета. Ограничения на выборку искомых документов при мониторинге проблемно-ориентированной информации могут быть разнообразны по тематике документов, территориальному или языковому признаку и т.д. Среди существующих подходов к мониторингу проблемно-ориентированной информации в сети Интернет можно выделить следующие:

Отбор документов путем самостоятельного исследования всего информационного массива сети Интернет посредством навигации и анализа документов на соответствие их заданной предметной области.

Отбор документов с использованием знаний об информационных ресурсах Интернета, которые предоставляют такие средства информационного поиска, как каталоги. Данный способ отбора информации достаточно прост и позволяет определить наиболее значимые ресурсы в заданной предметной области.

Отбор документов путем составления запросов существующим в сети Интернет средствам информационного поиска (информационно-поисковым поисковым системам и метапоисковым системам) с включением в них ключевых слов предметной области и последующего анализа отклика этих систем. Данный метод более эффективен, чем предыдущий, так как он осуществляет значительно больший охват информационных ресурсов Сети.

Другим важным этапом при формировании корпоративных знаний является **системный анализ** отобранной из сети Интернет информации.

Системный анализ – совокупность методов и средств исследования сложных, многоуровневых и многокомпонентных систем, объектов, процессов, опирающихся на комплексный подход, учет взаимосвязей и взаимодействий между элементами системы. **Системный анализ информационных ресурсов включает процессы индексирования, классификации и кластеризации.**

**Индексирование** – процесс выражения главного предмета или темы текста какого-либо документа в терминах информационно-поискового языка (знаковая система, предназначенная для описания основного смыслового содержания текстов (документов) или их частей, а также для выражения смыслового содержания информационных запросов с целью реализации информационного поиска).

Цель процесса индексирования – приписать некоторое множество идентификаторов (понятий, терминов, ключевых слов),отражающих содержание документа, т.е. составить поисковый образ документа (ПОД)

ПОД(Dm) = <t1,t2….tp>

Dm – текстовый документ, (m = 1, …, M; M – число текстовых документов);

ti – термин документа Dm, включаемый в поисковый образ ( ti принадлежит Tm; Tm–множество терминов документа Dm, i= 1, ..., p; p–число терминов документа, включаемых в поисковый образ документа).

Существуют различные типы систем индексирования:

К первому типу относят системы свободного индексирования. При этом способе из индексируемого документа выписываются в ПОД слова или словосочетания, которые отражают содержание индексируемого документа. Кроме этого, элементами ПОД могут быть слова, отсутствующие в этих элементах, но отражающие более точно смысл его текста с точки зрения средства поиска. Выписанные элементы упорядочиваются в алфавитном порядке. Такой упорядоченный набор слов (словосочетаний) представляет собой ПОД при этом типе индексирования.

При втором типе, который условно называют методом полусвободного индексирования, из документа выписывают слова и словосочетания вначале так же, как и при свободном индексировании. Однако выписанные элементы сравнивают затем с фиксированным словарем, не найденные в нем –устраняют, а оставшиеся, упорядоченные в алфавитном порядке, представляют собой ПОД.

Третий тип индексирования основан на статистическом подходе. Выбор слов (выражений) исходного текста, подлежащих включению в ПОД, производится на основе статистического анализа текста, при котором его слова рассматриваются как знаки, не имеющие семантических значений. При этом предлагались различные статистические критерии, основанные на сопоставлении относительной частоты употребления слова в документе, относительной частоты употребления слова в представительном массиве документов (т.е. в репрезентативной статической выборке) или распределение частоты слова по всем документам массива.

К четвертому типу относят системы индексирования, контролируемые заданным словарем (тезаурусом). Каждое слово текста сравнивается с точностью до основы со словарем, совпавшие слова записываются в ПОД.

Пятый тип, являющийся наиболее популярным в WWW в последние годы, основывается на ссылках на данный документ из других документов. В ПОД заносятся не только слова непосредственно из документа, выбранные одним из перечисленных способов, но и информация о ссылках на данный документ –число ссылок, текст этих ссылок и сведения об источнике информации, в котором они находятся.

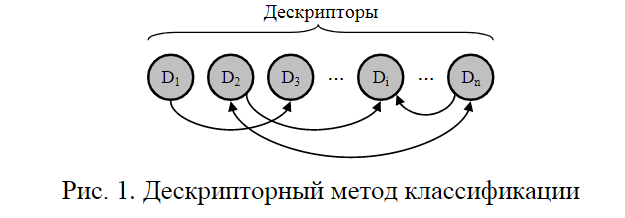
Следующий процесс системного анализа информации –классификация.

**Классификация**–система соподчиненных понятий (классов объектов) какой-либо области знания или деятельности человека, часто представляемая в виде различных по форме схем (таблиц) и используемая как средство для установления связей между этими понятиями или классами объектов, а также для точной ориентировки в многообразии понятий или соответствующих объектов. Классификация должна фиксировать закономерные связи между классами объектов с целью определения места объекта в системе, которое указывает на его свойства. Разработаны три метода классификации: иерархический, фасетный, дескрипторный. Эти методы различаются разной стратегией применения классификационных признаков.

Для организации поиска информации, ведения тезаурусов (словарей) эффективно используется дескрипторный (описательный) метод классификации, язык которого приближается к естественному языку описания информационных объектов. Особенно широко он применяется в библиотечной системе поиска системах поиска Интернета.

Суть дескрипторного метода классификации (рис. 1) заключается в следующем:

* отбирается совокупность ключевых слов или словосочетаний, описывающих определенную предметную область или совокупность однородных объектов, причем среди ключевых слов могут находиться синонимы;
* выбранные ключевые слова и словосочетания подвергаются нормализации, т.е. из совокупности синонимов выбирается один или несколько наиболее употребимых;
* cоздается словарь дескрипторов Di, т.е. словарь ключевых слов и словосочетаний, отобранных в результате процедуры нормализации.



Между дескрипторами устанавливаются связи, которые позволяют расширить область поиска информации. Связи могут быть трех видов:

* синонимические, указывающие некоторую совокупность ключевых слов как синонимы;
* родо-видовые, отражающие включение некоторого класса объектов в более представительный класс;
* ассоциативные, соединяющие дескрипторы, обладающие общими свойствами.

Достоинства дескрипторной системы классификации:

1)возможность создания большой емкости классификации;

2)сетевая структура, которая приводит к легкости внесения изменений

Недостатком дескрипторной системы классификации является сложность построения (при проектировании выполняется глубокий анализ предметной области).

**Кластеризация** –это автоматическое выявление групп семантически похожих документов среди заданного фиксированного множества документов. Группы формируются только на основе парной схожести описаний документов, характеристики этих групп заранее не заданы.

Кластеризация документов применяется для более эффективного хранения и выбора документов из коллекции. Она порождает однородные группы документов, которые более тесно связаны друг с другом, чем с документами из других групп.

При кластеризации информации имеется множество объектов D ={D1, D2, ..., Dn}, принадлежащих некоторой предметной области, каждый из которых характеризуется некоторым набором параметров T, представляющим вектор T ={t1, t2, ..., tk}.

Задача кластерного анализа заключается в том, чтобы на основе данных, содержащихся во множестве T, разбить множество объектов D на m (m-целое число) кластеров (подмножеств) S1, S2, ..., Sm так, чтобы каждый объект Dj принадлежал одному и только одному подмножеству разбиения и чтобы объекты, принадлежащие одному и тому же кластеру, были сходными, в то время как объекты, принадлежащие разным кластерам, были разнородными.

Значения близости (связей) между документами в коллекции несут информацию об их совокупной важности, релевантности по отношению к запросам коллекции. Это означает, что если некий документ отвечает поступившему в коллекцию запросу, то (в пространстве документов коллекции) в непосредственной близости от него могут быть найдены документы, в большей степени отвечающие документу запроса. Документы, не отвечающие данному запросу, не будут найдены вблизи отвечающих этому запросу документов. Важным при кластеризации является выбор меры близости объектов, от которого решающим образом зависит окончательный вариант разбиения объектов на группы при заданном алгоритме разбиения.

Меры близости, применяемые при кластеризации для многомерных пространств, разделяются на три типа:

-угловые меры (мера косинуса);

-меры, основанные на расстояниях;

-корреляционные меры;

Выделяются два основных класса методов кластеризации: разделяющие и иерархические. Основное их различие в структуре получаемых кластеров. Разделяющие на выходе предлагают набор выделенных в коллекции документов классов, причем, как правило, число классов должно быть задано заранее. Иерархические не имеют такого ограничения и строят вложенную иерархию кластеров. Однако такие методы имеют как серьезные ограничения в адекватности получаемых разбиений, так и в производительности.

Иерархические методы дают на выходе вложенную последовательность с единственным, включающим все остальные, кластером на вершине и простыми кластерами, не включающими в себя другие на самом низком уровне иерархии. Результат действия по таком алгоритму можно представить графически как дерево, именуемое дендрограммой, отражающее процесс слияния отдельных наблюдений в единый окончательный кластер. На нем наглядно представляется процесс объединения кластеров и получения промежуточных уровней дерева.

Такое представление является преимуществом данного метода, поскольку, помимо непосредственно объединения документа в классы, происходит разбиение документов на иерархически связанные классы, то есть сохраняется систематика набора документов.

Выделяют два основных подхода к созданию такой системы кластеров.

Агломеративный (накапливающий) подход. Начиная с элементарных документов, происходит их последовательное объединение в кластеры. Для этого подхода важно понятие сходства кластеров и расстояние между кластерами. В основном все предлагаемые алгоритмы отличаются стратегией объединения.

Дивизивный (делимый) подход. Начиная с одного, включающего в себя все документы кластера, происходит последовательное разбиение кластеров вплоть до получения элементарных документов. На каждом шаге необходимо принимать решение: какой кластер разделить и как произвести разделение. Дивизивными является большинство графоориентированных методов, построенных на последовательном разделении графа. Таким образом, кластерный анализ документов позволяет повысить полноту ответа на запрос, поскольку пользователь, нашедший документ, отвечающий запросу, может запросить и все документы, принадлежащие тому же кластеру. Повышается и точность поиска, поскольку в ответ на запрос будут возвращаться только документы, принадлежащие одному кластеру, объединенные в него на основе своей взаимной близости.

Основными проблемами большинства кластерных методов, причем не только иерархических, которые необходимо преодолевать для успешного выполнения нашей задачи, являются:

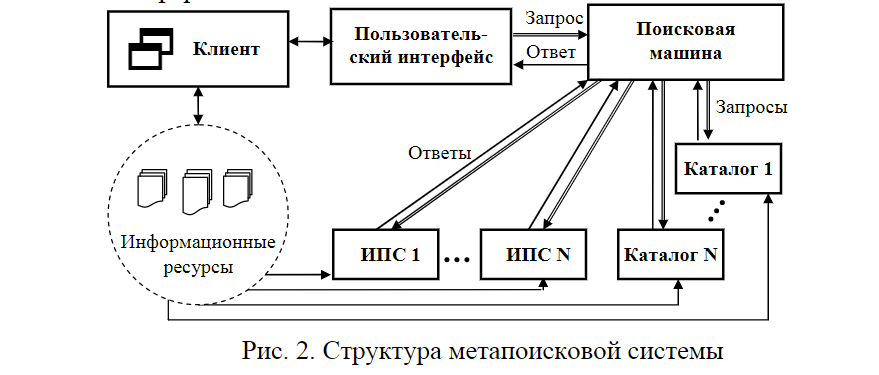
-большая размерность пространства;

-большой объем анализируемых данных;

-зависимость от вводимых параметров, определяющих результат анализа.

# Обзор существующих систем автоматизации мониторинга и системного анализа распределенной информации

Отличия в стратегии и широте охвата информационных ресурсов различных поисковых систем часто приводят к тому, что разные средства поиска дают разноречивые ответы на один и тот же запрос. Этой особенностью воспользовались разработчики метапоисковых систем, которые в своей работе используют потенциал других средств информационного поиска(рис.2).Алгоритм работы метапоисковых систем часто позволяет улучшить такие показатели,ткак «полнота» и «объективность» полученных результатов, а также сократить время, затрачиваемое на поиск информации.



Метапоисковые системы наиболее эффективны на начальных этапах поиска информации. Они позволяют быстро проверить, есть ли в Интернет необходимая информация и локализовать средства поиска, в которых она присутствует.

Web второго поколения—Web 2.0—предполагает существование единого информационного пространства в виде множества информационных единиц, которые могут распространяться по различным сайтам и сервисам.

Основные тенденции, характерные для среды Web 2.0:

- семантическая разметка информации и переход на XML;

-развитие web-сервисов и предоставление доступа к информации из любого места; возможность отчуждения информации от своего источника;

-независимая навигация и управление сайтом, позволяющие пользователям полностью контролировать интерфейс;

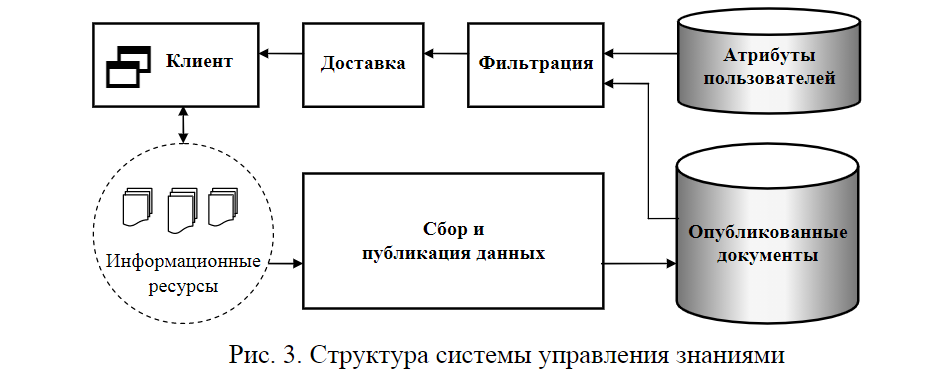
-отложенное добавление метаданных, осуществляемое сообществами пользователей;

-изменение парадигмы разработки, полное разделение структуры и дизайна сайта.

Однако, учитывая различия в подходах к обработке терминов разными системами, результат поиска (отклик системы) может быть не всегда релевантным запросу. В результатах поиска, как правило, присутствует информационный шум (не соответствующая требованиям пользователя информация).

Автоматизация управления знаниями в информационном пространстве осуществляется с использованием систем управления знаниями (knowledge management systems),которые предназначены для автоматического анализа и нахождения взаимосвязей между документами, людьми и информацией в пределах всей организации. По сути, это интегрированное решение, связывающее воедино все информационные источники организации и обеспечивающее прозрачный, централизованный доступ к ним, а также благоприятные условия для обмена и совместного использования данных.

Принципы работы **систем управления знаниями** (рис. 3.) состоят в следующем. Вся информация в системе либо публикуется непосредственно авторами материалов (поставщиками информации), либо автоматически копируется из внешних источников по заранее установленным правилам. Используя атрибуты опубликованных документов (эти атрибуты присваиваются документам авторами в момент публикации) и атрибуты пользователей, администратор системы может устанавливать персонификационные правила, позволяющие системе отбирать необходимое содержимое и показывать (или доставлять) его индивидуальным пользователям. Фильтрация содержимого может применяться при выдачах результатов поисковых запросов, формировании выборок, генерации персонифицированных каналов или страниц. Сформированные таким образом данные предоставляются пользователю посредством выбранного способа доставки: как веб-страница, сообщение электронной почты, канал веб-вещания.



Недостатки систем управления знаниями:

-используют на 80% гуманитарные технологии и только на 20% решения в области информационных технологий ;

-не позволяют автоматически находить новые источники знаний и выполнять их анализ без участия человека;

-не предоставляют возможности автоматически без участия человека формировать знания на основе информации из сети Интернет.

В результате проведенного анализа работ отечественных и зарубежных ученых, а также систем информационного поиска Интернет и систем управления знаниями были сделаны следующие выводы:

1. Ввиду наличия в Сети информации по всем областям знания использование ее в промышленных предприятиях и организациях является одной из наиболее актуальных проблем.

2. Для организации хранения и проведения системного анализа информационных ресурсов проанализированы технологии хранилищ данных(ХД), OLAPи DataMining, которые позволяют выполнять эффективный доступ информации.

3. Используемые методы для обработки информации Интернета, в том числе для поиска информации, и созданные с их применениемавтоматизированные программные системы реализуют лишь часть реально необходимых в организациях функций по управлению информацией.

К основным нереализуемым современными подходами и системами функциям относятся:

-описание проблемы в соответствии со структурой предметной области;

-автоматическое расширение накопленных знаний информацией из сети Интернет;

-автоматизированный анализ всего перечня результатов в средствах поиска;

-централизованное управление запросами и результатами поиска из различных систем.

# Анализ подходов к организации хранилища данных

Для лучшего осознания проблем реальной деятельности на предприятиях и в организациях создаются хранилища данных, которые позволяют расширить возможности оперативных систем. Несмотря на то, что хранилища данных бывают различных типов и могут опираться на разные методологии и даже философии, построения, все они имеют общие признаки:

1)Информация в хранилище данных организовывается вокруг базовых понятий, используемых в деятельности организации;

2)Исходные данные собираются из неинтегрированных оперативных и унаследованных приложений, очищаются от ошибок, затем агрегируются и представляются в виде, понятном конечным пользователям. Одними из основных новых возможностей, появляющихся в результате построения хранилищ данных, являются следующие:

- применение средств поддержки принятия решений на основе технологий интеллектуального анализа данных (Data Mining —добыча данных, knowledge discovery in databases —обнаружение знаний в базах данных), включающих методы логического вывода, нейронных сетей и нейрокомпьютеров, и др.

- использование средств, повышающих простоту поиска информации и обращения к конкретным прикладным функциям, например гипертекстовым, естественного языка, речевого ввода.

Одними из основных компонентов в архитектуре хранилища данных являются OLAP-сервер и OLAP-клиент (средство OLAP) [39]. OLAP-сервер отвечает за хранение данных, предназначенных для анализа, средство OLAP предоставляет пользователям возможность аналитической обработки этих данных.

Основным требованием к средствам OLAP является поддержка многомерного представления информации. На базе многомерного куба средство OLAP позволяет пользователям проводить анализ данных по определенной технологии. Таким образом, основной задачей OLAP-сервера является эффективное хранение многомерных кубов.

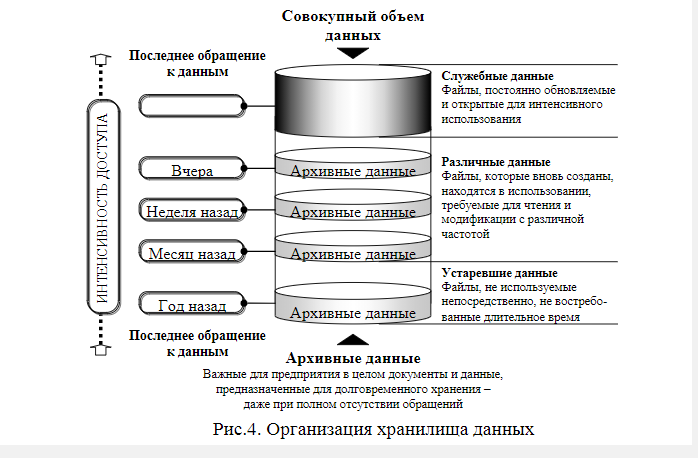
Хранилище информации представляет инструментарий и средства для доставки полной, точно, своевременной и понятной информации лицам, обладающим необходимыми полномочиями, а также для управления этой информацией с целью эффективной поддержки принятия решений.

В хранилище данных доступ к его содержимому и управление им требует соответствующих метаданных.

Вся информация о деятельности объекта исследований собрана в одном месте и доступна для исследования единым набором аналитических инструментов.

Структурно разработанное хранилище данных по предметной области можно представить в виде куба с измерениями: охват, время и потребность.

Созданное хранилище электронных ресурсов предприятия организовано с целью обеспечения гибкого централизованного администрирования хранения массивов информации, автоматизации управления жизненным циклом данных, увеличения надежности хранения и реализации возможности сокращения экономических затрат на программно-аппаратные решения, при работе с большими массивами данных (рис 4)

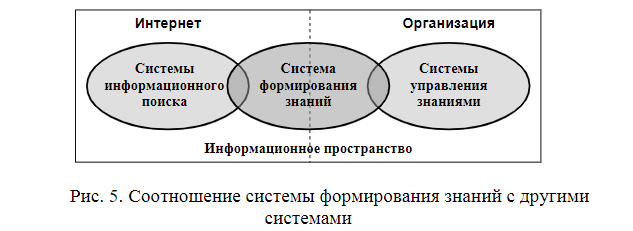


Таким образом, хранилище данных –это база данных, хранящая данные, агрегированные по многим измерениям. Данные из хранилища данных никогда не удаляются. Пополнение хранилища данных происходит на периодической основе. При этом автоматически формируются новые агрегаты данных, зависящие от старых.

# Система формирований знаний

Проведенный анализ существующих подходов к автоматизации мониторинга и системного анализа информации выявил значительный разрыв между системами, автоматизирующими эти процессы в Интернет (системы информационного поиска Интернет), и корпоративными системами, выполняющими автоматизацию их в отдельной организации (системы управления знаниями).

При этом процесс формирования знаний (получение знаний на основе моделей, методов и алгоритмов анализа данных) в существующих системах, автоматизирующих мониторинг и системный анализ, отсутствует. На основе изложенного сделан вывод о необходимости разработки систем нового типа –систем формирования знаний (рис. 5.), которые позволяли бы использовать значительно шире потенциал Интернет.



Рассмотренный тип систем –системы формирования знаний–позволяет эффективно выполнять мониторинг и системный анализ проблемно-ориентированной информации и таким образом способствует решению многих задач.

# Структура и обобщенный алгоритм работы системы формирования знаний

Структура системы формирования знаний состоит из следующих частей: модуль мониторинга, модуль индексирования, модуль аналитики, модуль поиска, модуль управления, база знаний и хранилище данных.

Одним из важных процессов в системе является мониторинг информации (реализуется модулем мониторинга). Для обеспечения качественного мониторинга проблемно-ориентированной информации в сети Интернет используются блоки базы знаний. Полнота охвата информационных ресурсов в системе достигается реализацией двух подходов, сущность которых состоит в следующем. Автоматизированная система периодически последовательно взаимодействует со всеми зарегистрированными, внешними по отношению к ней, средствами информационного поиска и таким образом отбирает проблемно-ориентированную информацию. Также в системе предусмотрена возможность задания экспертами адресов конкретных информационных ресурсов, соответствующих предметной области и также исследуемых системой.

Другим основным процессом представляемой системы является процесс индексирования информации(реализуется модулем индексирования).

На вход метода индексирования поступают документы, отобранные системой из внешней среды при помощи модуля мониторинга. На выходе метода индексирования получаются поисковые образы документов, которые необходимы системе формирования знаний для классификации, кластеризации информации, а также выбора документов, соответствующих запросу пользователя.

Процесс индексирования состоит из операций: отбор индексационных терминов, используемых для описания содержания документа, и приписывание этим терминам некоторого веса, отражающего их предполагаемую значимость. В модуле индексирования системы также выполняется процесс кластеризации информации, входе которого осуществляется предварительный автоматический анализ проблемно-ориентированной информации с целью создания фрагментов семантической сети, описывающих отобранные документы, и соотнесения их с основной семантической сетью системы формирования знаний, если она уже создана.

На заключительном этапе работы модуля индексирования документ помещается в хранилище данных, его поисковый образ–в блок метаданных базы знаний и информация о документе заносится в семантическую сеть системы.

Основным процессом в разработанной системе формирования знаний также является **классификация информации**.

В основе процесса классификации находится семантическая сеть базы знаний. Начальный этап формирования семантической сети осуществляется инженером по знаниям на основе знаний эксперта с использованием модуля управления. На этом этапе формируются узлы сети и им сопоставляются понятия предметной области. Затем осуществляется сопоставление узлам сети конкретных документов из заданной предметной области. Эта операция, как уже упоминалось, выполняется с использованием модуля индексирования. Разделение документов по узлам семантической сети может быть скорректировано в модуле управления с учетомзнаний эксперта.

Каждому узлу созданной в системе семантической сети может быть сопоставлено множество документов. Для упрощения анализа этих множеств документов в системе предусмотрен процесс кластеризации.

Решаемая задача –создание тематической структуры коллекции документов. Из данного на входе набора документов необходимо получить набор групп наиболее семантически близких документов, причем чтобы полученный набор групп имел иерархическую структуру. Основным инструментом решения задачи создания тематической структуры является кластерный анализ.

Иерархическая (вложенная) структура представляется в виде получаемых кластеров (классов). Это обусловлено необходимостью получения достаточно детализированной и систематической структуры коллекции. Требуется выделить не только общие классы, но и подклассы и так далее.

Получение кластеров связано с описанием различной формы, размера и плотности рассматриваемой информации. Вследствие неопределенности анализа кластеры могут быть различных форм: могут пересекаться, быть смешанными и так далее, различных размеров: тематики могут содержать документы в различном количестве, различной плотности. При выделении в коллекции несколько тематик одна из тематик может содержать набор тесно связанных, узкоспециальных документов, а другая –набор документов, специализированных по отдельности и отличающихся друг от друга в несколько большей степени, чем документы из первого набора.

Использование созданной семантической сети, которая классифицирует документы конкретной предметной области, конечными пользователями осуществляется через функцию аналитики модуля поиска. С использованием данного модуля пользователь формулирует запрос, который передается модулю поиска для отбора релевантной информации. Откликом системы на запрос являются фрагменты построенной семантической сети, которые можно рассматривать в качестве информационной модели изучаемой пользователем темы. Система позволяет осуществлять навигацию по выданным фрагментам сети и просматривать документы, которые им соответствуют.

Эффективность работы автоматизированной аналитической информационной системы в значительной степени определяется составом используемых в ней формализованных знаний. База знаний в системе–это связующее звено между хранилищем данных и модулями системы.

Основная составная часть базы знаний системы –блок метаданных. Это информация о том, что представляют собой данные хранилища, как они взаимосвязаны, какова структура данных, где хранятся данные, как получить доступ к ним и т.д. Доступ к метаданным имеют все модули системы, обслуживающие и использующие хранилище данных. Создание и поддержка блока метаданных производится администратором системы. Этот блок является основой для всех процессов в хранилище и для обеспечения конечных пользователей доступом к информации в нем. Следующий блок БЗ –семантическая сеть. Этот блок служит для описания предметной области и формализации знаний о группах документов хранилища данных, выделенных экспертом. Содержимое блока –семантическая сеть, узлы которой соответствуют наборам понятий и группам документов, относящихся к определенной теме, а дуги –отношения между ними. Семантическая сеть позволяет улучшить процесс поиска информации. Она классифицирует документы Интернета, относя их к определенным узлам, и обеспечивает значительное сужение диапазона поиска требуемых документов, что положительно сказывается на точности результатов поиска.

Блок базы знаний, имеющий название «блок лингвистического анализа», служит для понимания системой содержания документов и запросов пользователя на естественном языке. Он содержит знания, дающие возможность системе разбивать документы и запросы на отдельные слова и выделять из них словосочетания. Кроме этого, блок решает задачи выделения основ слов (слова без окончаний) и нахождения слов сходных по смыслу –синонимов.

Еще один блок базы знаний –блок когнитивного представления. Он служит для представления пользователю найденной с помощью данной автоматизированной системы информации в удобном для просмотра, восприятия и анализа виде и включает набор правил, необходимых для выполнения этих функций.