Новосибирский
государственный университет

Российская Академия Наук
Сибирское отделение
Институт вычислительных технологий

Некоторые проблемы построения фактографических информационных систем

В.Б.Барахнин, А.М.Федотов

Что такое Информация

«Многие вещи нам не понятны не потому, что наши понятия слабы: но потому, что сии вещи не входят в круг наших понятий»

Козьма Прутков

«Окружающий нас мир непознаваем, ввиду того, что мы изучаем не его, а лишь наше представление о нем»

Эммануил Кант

«Только теория решает, что можно наблюдать»

А.Эйнштейн

Проблема поиска информации

- одна из вечных проблем человеческого сообщества. На протяжении своего многотысячелетнего развития его представители неустанно находятся в поиске того, где находится что-либо: *пищи, жилища, пастбищ, дорог, сокровищ* и т. п.
- Человечество постоянно находится в поиске знаний: «информации о том, где лежат сокровища».

Проблема поиска информации

- Великий аргентинский писатель Хорхе Луис Борхес в своем эссе «Четыре цикла» писал, что в мировой литературе вечными являются четыре темы:
  - Падение города
  - Возвращение героя
  - Поиск
  - Самопожертвование бога.

Проблема поиска информации

- Нетрудно заметить, что наиболее часто встречающейся как в литературе, так и в реальности является третья тема – **поиск (информационная потребность)**, ибо четвертая тема выходит за рамки обычного человеческого опыта, а две первые проявляются лишь в «минуты мира роковые»

Можно выделить два типа информационных потребностей:

1. В сведениях об источниках необходимой научной информации
2. В самóй необходимой научной информации

Для удовлетворения информационных потребностей первого типа предназначены информационные системы, получившие название *документальных*, второго типа – *фактографических*.

Факты **- «особого рода предложения, фиксирующие эмпирическое знание» (Философская энциклопедия)**

Информационные технологии

Проблема доступа к информации является одной из основных проблем, возникающих в современной человеческой деятельности. Любой производственный или научный процесс порождает огромные объемы данных, и работать с ними становится все сложнее по мере того, как гигабайты данных превращаются в терабайты. Количество данных когда-нибудь превысит способность компьютеров их обрабатывать, поэтому необходимы новые технологии, инструментальные средства и алгоритмы для анализа этих данных.

Предназначение информационных систем

Если вам все равно, где вы находитесь, значит вы не заблудились

Организация хранение информации (организация хранилищ, поддержка систем хранения данных)

Управление информацией (добавление, модернизация, изменение данных)

Управление доступом к информации (контроль исполнения правил регламентации доступа к данным), идентификация данных

Поиск информации

Извлечение информации и предоставление ее пользователю в необходимом ему виде

Визуализация информации в соответствии требованиями пользователя

Предназначение информационных систем

Цель работы информационных систем – обеспечение конечного пользователя необходимой информацией.

Пользователя, как правило, не интересует, как устроена технологическая «кухня» информационной системы.

Более того, чем меньше эта «кухня» пользователю заметна, тем лучше построена та или иная информационная система.

Следует думать, что использование проверенных временем и практикой типовых решений позволит обеспечить эту технологическую «прозрачность».

ПРОБЛЕМЫ

Сложные проблемы всегда имеют простые, легкие для понимания неправильные решения

Закон Х.Л.Менкина

модели и стандарты представления информации и метаинформации;

автоматическая классификация информации;

доступ к распределенным и разнородным коллекциям (интероперабельность, масштабируемость, обнаружение релевантной информации, интеграция метаинформации);

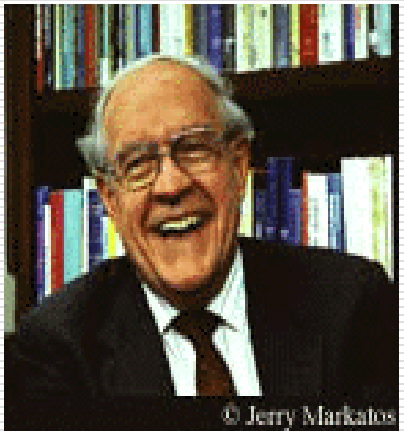
интерфейсы пользователей, визуализация и анализ данных,

вопросы интеллектуальной собственности;

Требования

- Требовать и эффективности, и гибкости от одной и той же программы - все равно, что искать очаровательную и скромную жену... по-видимому, нам следует остановиться на чем-то одном из двух.

- *Фредерик Брукс-младший*



Информационные технологии

Кабы схемку иль чертеж,
Мы б затеяли вертеж,
Ну а так - ищи сколь хочешь,
Черта лысого найдешь!

Л. А. Филатов. «Про Федота-стрельца, удалого молодца»

Вместе с тем предъявляются серьезные требования к обеспечению прозрачного доступа и долговременной сохранности «информации».

А в результате вопросы «что хранить?», «как хранить?» и «как найти?» остаются самыми существенными: без ответа на них все остальные теряют актуальность.

Необходима технология извлечения фактографической информации из научных документов достаточно произвольной структуры.

Что такое «ФАКТ»
Понятие «факт» - ГОСТы

В официальных документах ГОСТ 7.73–96 «Поиск и распространение информации» и ГОСТ 7.74–96 «Информационно-поисковые языки» этот термин «*факт*» практически не формализован.

Так, в ГОСТ 7.74–96 дано лишь косвенное, причем не слишком содержательное, определение факта: «7.7. **фактографическое индексирование:** Индексирование, предусматривающее отражение в поисковом образе документа конкретных сведений (фактов)».

Если бы комплименты были правдой, это были бы не комплименты, а информация.

Кретя Патачкувна "Моя кибернетика», в книге «Мысли людей великих, средних и пса Фафика»

Уточнение понятия «факт» - энциклопедии и монографии:

1. Факты следует отличать от *данных*, фиксирующих специфику объекта, условия наблюдения и т. п. Понятие же научного факта «предполагает элиминирование такой информации, т. е. требует определенного *обобщения* непосредственных данных».

2. Фактом можно назвать лишь знание, выдержавшее критическую проверку, то есть полученное в результате обобщения и переработки данных абстрактно-логическим мышлением.

Когда мы пытаемся вытащить что-нибудь одно, оказывается, что оно связано со всем остальным.

Закон Муира

Уточнение понятия «факт» - энциклопедии и монографии:

3. Любой факт, прежде чем стать объектом научной коммуникации, должен быть преобразован (***закодирован***) в текст или изображение, получив форму научного документа или его части. Более того, «объектом сбора, хранения, поиска и выдачи в так называемых фактографических информационно-поисковых системах… могут быть лишь соответствующие тексты или документы, описывающие некоторые данные или факты, если под документом понимать… любой фрагмент такого текста» .

Уточнение понятия «факт» - монография ВИНИТИ

Тенденция стирания граней между понятиями «данные» и «факты» отчетливо проявилась в более современной монографии ВИНИТИ (1996).

Данные понимаются как факты и идеи, представленные в символьной форме, позволяющей проводить их передачу, обработку и интерпретацию, а *информация* – как смысл, приписываемый данным на основании известных правил представления фактов и идей.

Структурированная (связанная причинно-следственными и иными отношениями) информация, образующая систему, составляет *знания*.

Уточнение понятия «факт» - семантический подход

- Семантический подход к информации прагматичен, он позволяет провести разделение между данными и информацией.
- Данные — это представление фактов и понятий в форме, пригодной для их передачи и интерпретации, а
- Информация — это смысл, который ЧЕЛОВЕК приписывает данным на основании известных ему правил их представления (моделей).

Уточнение понятия «факт» - семантический подход

informatics - научное направление, изучающее модели, методы и средства сбора, хранения, обработки и передачи информации - совокупность дисциплин естественно объединяющихся с целью семантической (смысловой) обработки информации

Эдсгер В. Дейкстра

Информатика не более наука о компьютерах, чем астрономия — наука о телескопах.

Уточнение понятия «факт» - семиотический подход

На основании семиотического подхода нами было показано, что данные соответствуют синтаксическому уровню сообщения (в том числе документа), информация (в узком смысле!) – семантическому, а знания – прагматическому.

Отсюда вытекает, что функционирование интеллектуальной информационной системы основано на двух противоположных процессах: при пополнении ИнтС новыми сведениями происходит преобразование семантической информации в данные, однако непосредственно потребности пользователя удовлетворяет обратный процесс – извлечение из данных нужной пользователю информации и знаний.

Уточнение понятия факт - связь с моделью

Итак, в качестве «первичного» факта (т.е. содержащегося непосредственно в тексте документа) рассматривается некоторая семантическая информация, но в справочно-информационный фонд ИнтС факт заносится в качестве элемента данных, что соответствует уже упоминавшемуся соотнесению данных и фактов.

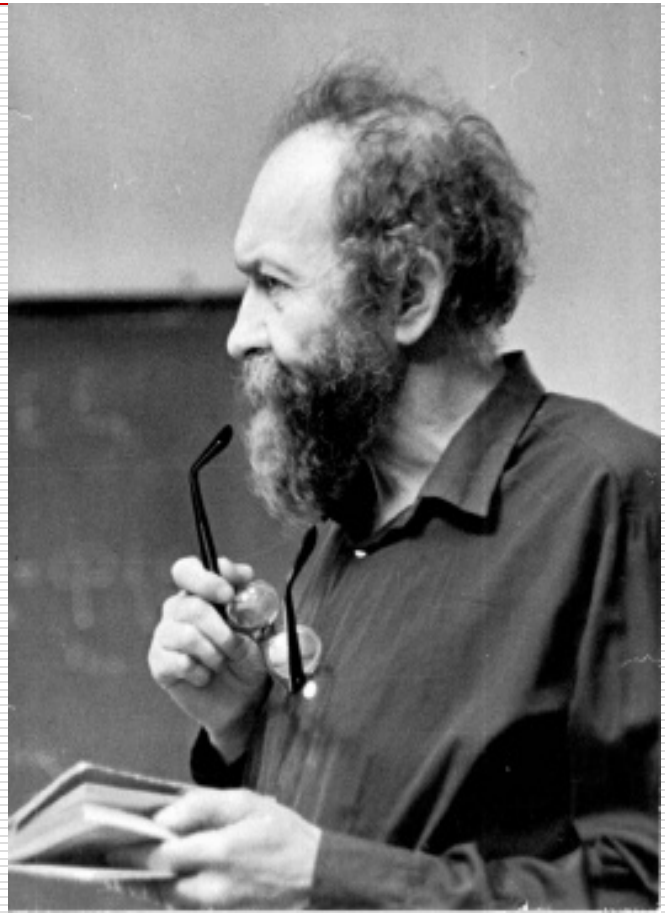
Но какого рода информация может быть занесена в справочно-информационный фонд системы в виде данных? Ведь сами по себе данные не несут никакой информационной ценности без соответствующих моделей: ***«Нет модели – нет информации»*** (А.А.Ляпунов)

В качестве модели предметной области обычно выступает ее *онтология* (какая именно смысл мы вкладываем в это весьма широко трактуемое понятие – будет уточнено в следующем разделе).

Информационные модели

- Как неоднократно отмечал А.А.Ляпунов: ***"нет модели - нет информации"***.

- Перефразируя А.А.Ляпунова следует отметить, что ***"конечная цель всей работы, связанной с применением информационных технологий - является понимание того или иного явления, а не получение каких-либо чисел или картинок"***.



Информационные модели

– Скажите, пожалуйста, куда мне отсюда идти?

– Это во многом зависит от того, куда ты хочешь прийти, – ответил Кот.

– Да мне почти всё равно, – начала Алиса.

– Тогда всё равно, куда идти, – сказал Кот.

– Лишь бы попасть куда-нибудь, – пояснила Алиса.

– Не беспокойся, куда-нибудь ты обязательно попадёшь, – сказал Кот,

– конечно, если не остановишься на полпути.

Льюис Керролл, «Алиса в стране чудес»

Используемое определение факта

При создании фактографических информационных систем разумно следующее понимание факта:

***входящая в текст документа характеристика сущности, описываемой в справочнике (онтологии) информационной системы, представляемая как единичное значение данных*.**

Именно онтология фактографической системы определяет, чтó будет считаться фактом в рамках этой системы. Здесь мы имеем дело с ситуацией, столь характерной для естественных наук, о которой говорил, например, А.Эйнштейн в своей известной беседой с В.Гейзенбергом: ***«Только теория решает, что можно наблюдать»***

Особенности онтологий для фактографических систем

Под онтологией понимают широкий спектр структур, представляющих знания о той или иной предметной области с разной степенью формализации:

1. словарь с определениями;
2. простая таксономия;
3. тезаурус (таксономия с терминами);
4. модель с произвольным набором отношений;
5. таксономия и произвольный набор отношений;
6. полностью аксиоматизированная теория.

Особенности онтологий для фактографических систем

Тезаурус становится онтологией тогда, когда связи между дескрипторами не просто эксплицированы (как это предусмотрено в классическом определении тезауруса), но и классифицированы универсальными зависимостями (правилами) типа «общее – частное», «часть – целое», «причина – следствие» и т.п.

Разумеется, это – лишь «нижняя граница» сложности онтологии. Для эффективной работы с фактами следует, чтобы сущности, относящиеся к предметной области, были представлены не только обозначающими их терминами, но и достаточно широким набором атрибутов, т.е. речь идет об онтологии, обладающей известными признаками модели предметной области.

«Есть правила для выбора решения, но нет правил для выбора этих правил»

Энон

Онтология как модель предметной области

В роли онтологии – модели предметной области – может выступать та или иная модель интеллектуальной информационной системы информационной системы, например, имеющая вид:

$$S = \langle K, M, M_j \langle K_i, K_i' \rangle \rangle,$$

где K — классы сущностей, M — множество используемых атрибутов сущностей, $M_j \langle K_i, K_i' \rangle$— типы возможных связей между классами сущностей, когда сущность из класса K_i' может входить в качестве значения атрибута M_j сущности из класса K_i.

Онтология как модель предметной области

Тем самым любая сущность представляется как множество значений атрибутов сущности (с учетом возможных повторений)

При создании информационной системы сущности будут представлены в виде описывающих их документов, а атрибуты сущностей будут представлять собой элементы метаданных.

Возможность расширения модели онтологии

Модель может быть расширена и усложнена, например, путем явного введения необходимых ограничений на атрибуты сущностей (которые можно назвать иначе: *аргументы фактов*) с использования структуры, называемой *схемы фактов* (*Загорулько Ю.А., Сидорова Е.А., Боровикова О.И.)*

Возможность расширения модели онтологии

Схема факта *Fk* представляется в виде ‹*Arg*, *Rs*, *C*›, где

Arg – множество дескрипторов аргументов факта (в качестве дескриптора может выступать быть тип словарной единицы, класс информационного объекта и т.п.);

Rs = ‹*t*, *op(t)*, *P*› – результат применения схемы, где *t* задает класс объекта, *op(t)* – тип операции, применяемой при условии выполнения ограничения

C , *P* – множество правил для формирования или редактирования объекта, каждое из которых ставит в соответствие атрибуту результирующего объекта либо точное значение атрибута, либо значение атрибута одного из аргументов; наконец, *C* – множество ограничений (морфологических, синтаксических, семантических, структурно-текстовых), накладываемых на характеристики аргументов факта.

Автоматизированное извлечение фактов из документов

«…не существует сколько-нибудь значительных различий в теории и методике построения документальных и фактографических информационно- поисковых систем, если фактографический поиск понимать лишь как процесс отыскания уже готовых данных и фактов, ранее введенных в фактографическую систему… Однако под фактографическим поиском можно понимать и нечто принципиально иное, а именно отыскание машиной требуемых данных и фактов в текстов научных документов, написанных на одном или нескольких разных естественных языках, …[что] требует оперирования со смыслом текстов его анализа и синтеза, т.е. моделирования достаточно сложных мыслительных процессов» (ВИНИТИ, 1976)

Автоматизированное извлечение фактов из документов – табличные данные

Табличные данные могут выступать в качестве фактов, если являются, например, характеристиками предметов, географических объектов и т.п. Для их извлечения из документов существуют разнообразные, весьма надежные алгоритмы.

Автоматизированное извлечение фактов из документов – массивы однородных слабоструктурированных текстовых документов

При занесении фактов, содержащиеся в массивах однородных документов, описывающих предметную область: биографических справочниках, геологических, ботанических или зоологических каталогах и т.п., наиболее целесообразно использовать алгоритмы, учитывающие информацию о закономерностях их текстовой структуры (например, общих для всех документов массива синтаксических и семантических конструкциях), а также о гипертекстовой разметке обрабатываемых документов (при наличии таковой).

Такой алгоритм, извлекающий факты (метаданные) о библиографии документов, подробно описан, например, в монографии Ю.И.Шокина, А.М.Федотова, В.Б.Барахнина.

Автоматизированное извлечение фактов из документов – массивы однородных слабоструктурированных текстовых документов

Он может быть легко адаптирован к фактографической информации произвольного характера, содержащейся в массивах документах, имеющих более или менее однородную текстовую структуру.

Автоматизированное извлечение фактов из документов – тексты произвольного характера

Задача извлечения фактов из произвольных текстов на естественном языке до сих пор, по-видимому, не имеет сколько-нибудь общего решения, поскольку построение такого решения предполагает, в частности, достаточно точное моделирование когнитивной деятельности человека, а также наличие мощных средств как синтаксического, так и семантического анализа текстов, включая подробнейшие онтологии, тезаурусы которых учитывают, например, всё богатство синонимии естественного языка (не столько даже в части научной лексики, сколько в части лексики общеупотребительной).

Автоматизированное извлечение фактов из документов – тексты произвольного характера

«Частное решение» этой задачи применительно к той или иной предметной области предполагает, прежде всего, построение онтологии, тезаурус которой включает, наряду с описанием сущностей предметной области, по крайней мере, те пласты общеупотребительной лексики (разумеется, с учетом синонимии), которые наиболее характерны для данной области.

Непосредственная работа по извлечению фактов из текста может опираться на совокупном применении методов синтаксического и семантического анализа.

О взаимодействии фактографических систем с пользователями

Нередко в качестве чуть ли непременного атрибута качественной фактографической системы называют возможность формулировки запроса на естественном языке.

Однако задача «понимания» системой запросов на естественном языке в каком-то смысле эквивалентна задаче извлечения фактов из текстов на естественном языке. При этом следует учесть, что далеко не все пользователи способны формулировать свой вопрос так четко и недвусмысленно, как это умел делать на своем знаменитом семинаре в Институте физических проблем АН СССР Л.Д.Ландау:

С первых слов, как Вельзевул во плоти,
Навалился Дау на него:
«Лучше вы скажите, ***что в работе***
Ищется как функция чего?»

О взаимодействии фактографических систем с пользователями

Слишком же расплывчатая постановка вопроса, «не распознанная» информационной системой, может привести к тому, что у пользователя сложится ошибочное мнение, будто бы система не располагает необходимой ему информацией.

Коль скоро мы рассматриваем в качестве фактов характеристики сущностей, описанных в онтологии, то весьма несложный интерфейс, позволяющий просматривать онтологию посредством использования последовательности гиперссылок (или даже посредством таблицы), сможет предоставить пользователю возможность без труда найти нужный факт или, по крайней мере, убедиться в том, что этот факт не занесен в систему.

О взаимодействии фактографических систем с пользователями

Еще одним важным примером практического использования фактографических систем в научных издательствах и редакциях журналов может служить проверка достоверности сведений, содержащихся в рукописях, имеющих биографический, научно-публицистический, обзорный и т.п. характер. Факты, извлекаемые из текста рукописей, подвергаются сравнению с «эталонными» фактами из онтологии информационной системы, и в случае расхождения редакция просит автора уточнить правильность приведенных им сведений.

**Большинство людей не получают того, чего хочет,
а всё потому, что сами не знают чего хотят.
Как исполнять желания, если их нет?
Дж.Б.Пристли. Тридцать первое июня.**

Заключение

В данной работе намечены основные шаги в направлении разработки технологии обработки фактографической информации, содержащейся в научных документах достаточно произвольной структуры. Показано, что при создании фактографических информационных целесообразно следующее понимание факта: **входящая в текст документа характеристика сущности, описываемой в онтологии информационной системы, представляемая как единичное значение данных**. Предложена простейшая модель онтологии фактографической системы.

Важным этапом практической реализации предлагаемых в статье подходов должна стать реализация алгоритмов синтаксического и семантического анализа текстов с целью извлечения фактов.