Лекция 12. Информационный поиск как приложение БЛП

Информация – это совокупность сведений, которые информационная система воспринимает из окружающей среды, выдает в окружающую среду либо сохраняет внутри себя.

Информационный поиск – это некоторая последовательность операций, выполняемых для отыскания документов (статей, научно-технических отчетов, книг и т.д.) и выдачи фактических данных, представляющих собой ответы на заданные запросы.

Информационный поиск производится при помощи систем информационного поиска как специальных комплексов программного, информационного и технического обеспечения. СИП должна обеспечивать выполнение следующих функций:

– Приём информации и её предварительная обработка.
– Анализ документов и данных, а также, возможно, хранение.
– Анализ и организация информационных запросов.
– Поиск релевантной информации.
– Выдача запрашиваемой информации.

Задача информационного поиска

В общем виде задачу информационного поиска можно сформулировать следующим образом: по некоторой заданной информации J_1 необходимо найти и выделить определённые части J_3 информации J_2. Таким образом, в результате некоторого взаимодействия F информации J_1 и J_2 выделяют информацию J_3.

$$J_1 = F\,(J_1, J_2)$$

Реализацию оператора F в общем случае и называют информационно-поисковой системой.

Выполнение основных функций ИПС обеспечивается следующими её структурными элементами: информационно-поисковым языком, базой данных индекса, поисковой машиной и другими техническими средствами.

Информационно-поисковый язык представляет собой формализованную знаковую систему, предназначенную для выражения основного смыслового содержания документов и запросов с целью отыскания в массиве тех документов, которые отвечают на поставленный запрос.

Поисковый образ документа – это выраженное в терминах ИПЯ основное смысловое содержание документа, которое поставлено в однозначное соответствие этому документу и по которому отыскивается данный документ в массиве документов.

В информационных системах формализации подвергаются не только первичные документы, но и информационные запросы. Выраженное в терминах информационно-поискового языка основное смысловое содержание запроса называется *поисковым предписанием* или *поисковым образом запроса***.**

Процедуру выражения основного смыслового содержания документов и

запросов в терминах ИПЯ обычно называют *индексированием***.**

В настоящее время индексирование выполняется как вручную, так и автоматически. Некоторые СИП в Интернет используют для этих целей специальные программы (роботы-индексировщики), которые сканируют сеть на предмет поиска новых документов и поддержания базы данных индекса в актуальном состоянии.

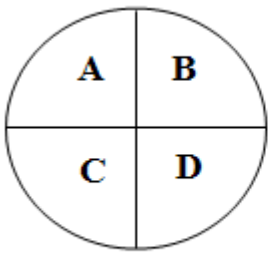
Ручной процесс индексирования (в библиотеках) состоял в том, что каждому документу приписывался набор терминов, которые он описывает, таким образом, с каждым документом связывался возможный к нему запрос. Автоматическое построение индекса предполагает, что текст запроса и документа будут обработаны одной процедурой обработки, а затем в численном виде процедура сравнения определит релевантность текста документа тексту запроса.

Вопрос о выдаче или невыдаче документа решается на основании критерия смыслового соответствия (релевантности), т.е. правила, на основании которого выясняется степень смысловой близости поискового образа документа и поискового образа запроса.

Термин *релевантность* – это один из основных терминов в теории информационного поиска. Релевантность – это степень соответствия документа или части текста документа запросу.

Целью систем информационного поиска является получение как можно большего числа релевантных документов при как можно меньшем проценте нерелевантных.

Основными параметрами, характеризующими эффективность работы ИПС, являются *коэффициент полноты* и *коэффициент точности* (см. рисунок 1)**.** Первый определяется отношением количества выданных релевантных документов к общему количеству релевантных документов в базе данных индекса. Второй – отношением количества выданных релевантных документов к общему количеству выданных документов.



A, B – найденные документы
A, C – релевантные документы

$$\text{Точность} = \frac{A}{A+B}, \quad \text{Полнота} = \frac{A}{A+C}$$

Рисунок 1.

Общая схема функционирования документальной СИП представлена на Рисунке 2.

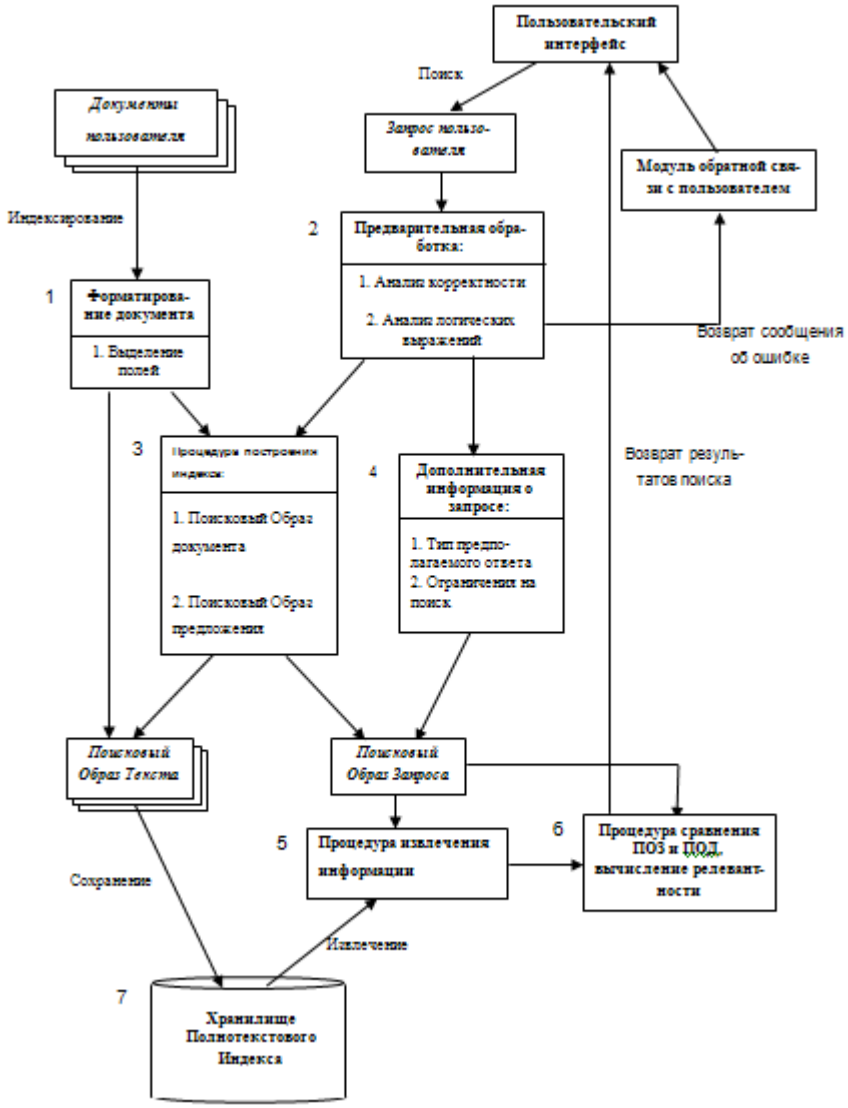


Рисунок 2 – Общая схема функционирования СИП.

Схема СИП, представленная на рисунке 2 состоит из следующих основных компонетов:

1. **Преформатор** документов. Основной задачей преформатора является преобразование документов пользователя в унифицированный формат (например, XML) и выделение значимых полей документа (для статей это могут быть: “Название”, “Аннотация” (Abstract), “Основное тело”, “Библиографические ссылки”, “Автор” и т.д.).

2. **Предварительная обработка запроса пользователя**. Данный модуль анализирует запрос пользователя на принадлежность языку запросов, может проверить грамматическую правильность слов, произвести построение логического дерева запроса (запрос может включать в себя как логические операторы, так и фразы, предложения естественного языка).

3. **Процедура построения индекса**. Данная процедура сильно отличается в разных системах. Конкретная реализация зависит от потребностей пользователя, т.е. от типа запросов и характера документов. На максимальном уровне понимания текста она состоит из следующих шагов:

- Выделение границ слов и предложений – грамматическое форматирование.

- Лексический анализ – определение частей речи и построение словоформ.

- Грамматический анализ – построение дерева грамматических зависимостей между словами.

- Семантический анализ – выделение фактов, знаний, семантический анализ имен ных групп, выделение причинно-следственных отношений.

4. **Выделение дополнительной информации о запросе.** Этот модуль может определить тип предполагаемого ответа, удалить неинформативные слова, фразы, произвести дополнительный анализ параллельно процедуре индексирования для увеличения полноты и проч.

5. **Процедура извлечения информации основанная, на ПОЗ.** Данный модуль по заданному запросу получает из хранилища документы (предложения или фразы). Процедура может быть реализована по-разному, например, уже на данном этапе можно извлекать документы с определенной релевантностью и возвращать их пользователю. С другой стороны, в целях экономии места, используемого для хранения индекса, данная процедура может возвращать приблизительно релевантные документы, а процедура 6 будет дополнительно их сравнивать с ПОЗ и извлекать более точные ответы.

6. **Процедура сравнения ПОЗ и ПОД, вычисление релевантности.** Данная процедура определяет близость ПОД ПОЗ в численном виде. В большинстве случаев требуется процедура нечеткого сравнения с заданной функцией штрафов для отказов.

7. **Хранилище Полнотекстового Индекса.** В зависимости от языка запросов пользователя, глубины анализа текста, количества и объема документов, а также потребностей пользователя в обновлении хранимой информации индекс может быть представлен по разному (см. Способы представления индекса).

По виду выдаваемой информации ИПС можно разделить на *документальные* и *фактографические*. Наряду со многими общими чертами эти системы имеют существенные различия: разные цели поиска, разные формы составления запросов и хранения информации, различия в точности поиска.

В *документальных системах* для составления запроса чаще всего используют набор слов из текста документа, его реферата, аннотации или библиографического описания. В ответ на запросы такие системы выдают подборки документов, кроме того, эти системы допускают поисковый шум, т.е. нетождественность поисковых предписаний информационных запросов поисковым образам документов. Документальные ИПС широко используются в библиотеках, архивах, для поиска документов в сети Интернет и т.д.

Фактографические (или *вопросно-ответные*) *системы* используют для нахождения некоторых конкретных фактов, т.е. поисковый шум не допускается.

Как правило, обычные системы информационного поиска в качестве ответов возвращают список релевантных документов, далее пользователь работает с этими документами сам, т.е., как описывалось выше, в “идеале” он должен их все просмотреть. *Вопросно-ответные системы* в качестве результата возвращают только ту часть текста документа, которая непосредственно отвечает на заданный вопрос (запрос). Если же вопрос

поставлен таким образом, что ответом на него является целый документ, то вопросно-ответная система должна вернуть запросно-ориентированный реферат данного документа со ссылкой на оригинальный документ. Схематически простейшую вопросно-ответную систему можно представить так, как изображено на рисунке 3.

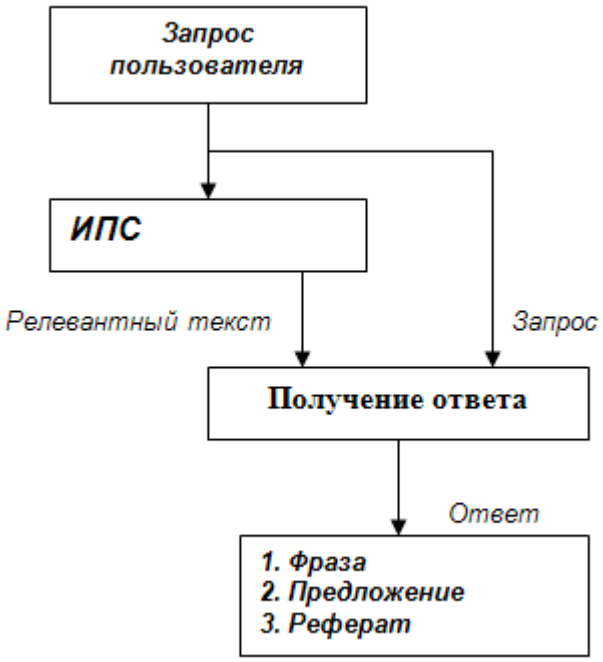


Рисунок 3 – Простейшая вопросно-ответная система.

Их разработка усложняется также тем, что ответы на вопросы типа: «В каком году была создана ООН?» или «Кто был президентом США в 1988 году?» требуют очень точного понимания смысла вопроса, а также довольно сложных моделей хранения информации. Использование формальных грамматик для разбора запроса пользователя, а также форматированного представления данных сильно сужает область применения таких систем, т.к. на сегодняшний день пока ещё не разработан метод, который даст компьютеру возможность полностью понять любую фразу естественного языка. Поэтому такие системы обычно принимают запросы, составленные по некоторому шаблону и относящиеся к какой-то определённой предметной области.

В настоящее время основным направлением развития фактографических систем является разработка методов выдачи не только информации, уже хранящейся в системе, но и полученной в результате переработки имеющихся данных. Повышенный интерес исследователи проявляют также к разработке методов понимания запросов на естественном языке.

Однако в общем случае задача точного ответа на вопрос еще не решена. Можно говорить лишь о решении для определенного класса вопросов с определенной точностью и полнотой.

Информационно-поисковые языки

Типы и виды ИПЯ

Способ задания лексических единиц

1. *Контролируемые* — языки, словарный состав которых задается и контролируется с помощью словарей и таблиц. К ним относят различные системы классификации (УДК, ББК, классификация Дьюи).
    1. Язык предметных рубрик. На основе иерархической классификации строят систематические каталоги. На основе языка предметных рубрик строят предметные каталоги. Алфавитные каталоги — ручной поиск.
    2. Дескрипторные ИПЯ.
2. *Неконтролируемые* — лексика не задается словарем, а строится на основе выбора терминов естественного языка. Такие ИПЯ широко начали применяться в последнее время.

Порядок записи лексических единиц

1. Некоординируемые языки — не допускающие координации своих лексических единиц (нет связи между ними) ни в процессе индексирования, ни в процессе поиска. (система расстановки книг в библиотечном фонде, по инвентарным номерам).
2. Координируемые ИПЯ — языки, в которых лексические единицы связывается, координируются между собой или в процессе индексирования или в процессе использования.
    1. Предкоординируемые — связи между лексическими единицами устанавливаются перед поиском.
    ИПЯ , построенные на принципах предкоординации. Эти языки представлены известными классификационными системами вида:УДК, МКИ, ГРНТИ, ББК, ТБК и др.
    Предкоординация [pre-coordinatin] - Построение словарного состава ИПЯ (до его использования при индексировании), которое характеризуется применением словосочетаний и фраз, выражающих сложные понятия.
    2. Посткоординируемые — когда связи между лексическими единицами устанавливаются только при поиске.
    Посткоординация [post-coordination] - Построение словарного состава путем разделения сложных понятий на составные элементы и последующего объединения полученных **лексических единиц** ИПЯ при **индексировании** документов вводимых в информационно-поисковые массивы и запросов путем использования **логических операторов** и других средств, представляющих его **синтаксис .**

Главными компонентами информационно-поискового языка являются словарный состав и синтаксис.

Словарный состав – это совокупность слов, используемых в информационном языке. Словарный состав естественного языка – это все его слова.

Синтаксис языка – это совокупность тех правил, согласно которым из элементов словарного состава строятся фразеологические единицы со значениями, которые невозможно выразить отдельными словами из основного словарного состава. Примером синтаксической единицы в ЕЯ служит обычное предложение.

Для сравнения различных ИПЯ используют следующие атрибуты: семантическая сила, многозначность, компактность.

Семантическая сила – это способность языка идентифицировать предметы, различать их мелкие особенности и давать описание предмета с различной степенью детализации. Очевидно, что наибольшей семантической силой обладает ЕЯ. Большая степень точности описания предмета при использовании языков с большой семантической силой достигается как за счёт большого словарного запаса, так и благодаря наличию широкой возможности использования на основании синтаксиса различных комбинаций дескрипторов.

Многозначность для ИПЯ означает, что слово или синтаксическая единица может иметь более одного значения, а также некоторое отдельно взятое значение может иметь более одного символического представления в словарном запасе языка. Наиболее многозначными являются ИПЯ с богатым синтаксисом. С одной стороны, многозначность затрудняет организацию процесса поиска, а с другой – открывает широкие возможности выбора при описании запросов и документов.

Компактность характеризует физический размер, иначе, длину индексирующего термина или поискового образа, которая требуется для передачи определённого количества информации. Вопрос компактности особенно важен, если СИП базируется на множествах огромного объёма. Среди наиболее употребительных ИПЯ естественный язык выделяется как малокомпактный.

Основные типы поиска

1 Поиск по ключевым словам

Выделяют список ключевых слов. Под *ключевым словом* понимается определенная лексическая единица, являющаяся либо отдельным словом, либо словосочетанием естественного языка. Основным критерием выбора таких слов является степень их полезности или эффективности в документах и запросах. Например, для выбора 15-20 тыс. ключевых слов необходимо обработать документы объемом 100 тыс. слов. При этом учитывается частота встречаемости, частота сочетаемости и т.д.

На базе созданного массива ключевых слов формируется дескрипторный словарь. *Дескриптор* – это ключевое слово, выбранное из группы условно-

эквивалентных ключевых слов и представляющее данную группу при индексировании. Из 15-20 тыс. ключевых слов обычно получают около 6 тыс. дескрипторов.

Простые дескрипторные словари без грамматики обычно разрабатываются для узкой области и обеспечивают высококачественный поиск информации. Если же ИПС рассчитана на справочно-информационные фонды большого объема и с широкой тематикой, то точность поиска можно обеспечить только за счет увеличения количества дескрипторов в ПОДах, т.е. за счет глубины индексирования. Однако при этом растет количество ложных сочетаний, которые можно составить из дескрипторов, входящих в один и тот же поисковый образ, т.е. возникает поисковый шум. Поэтому при поиске информации в таких системах используют более сложные дескрипторные языки, обладающие сложными грамматическими средствами, так называемыми информационно-поисковыми тезаурусами.

Под *тезаурусом* понимается словарь-справочник, в котором перечислены дескрипторы и синонимичные им ключевые слова, показан метод устранения многозначности и выражены родовидовые и ассоциативные связи между дескрипторами. Такая система позволяет быстро найти ту группу дескрипторов, которая связана с интересующим индексатора вопросом.

2 Поиск по запросу булевского типа

Наиболее распространенным ИПЯ среди поисковых систем сети Интернет является *ИПЯ булевого типа*,

Под логическим запросом понимается запрос, содержащий в себе логические или контекстные операторы с целью уточнения, сужения или расширения области поиска. Такой запрос может содержать слова, разделенные пробелами, операторами или кавычками и операторы, позволяющие составить логические выражения из терминов, связанных логическими операторами AND, OR, NOT (И, ИЛИ, НЕ).

Например, запрос *((информационная and система) or ИПС) not СУБД* означает, что нужно найти все документы, которые содержат либо слова *информационная* и *система* одновременно, либо слово *ИПС*, но при этом не содержат слово *СУБД*. Терминами в запросах *булевого типа,* как правило, выступают слова и фразы, которые требуют точного совпадения.

В таблице приведены поддерживаемые операторы в порядке возрастания их приоритета:

Таблица – Список операторов запросов и их синтаксис

Оператор	Описание	Синтаксис	Тип
<OR>	Документ должен содержать какое-либо слово-операнд	$Слово_1$ $<оператор_1>$ $слово_2$ $<оператор_2>$ $слово_3$…	Бинарный
<NEAR>	Слова-операнды должны находиться в пределах одного предложения		
<AND>	Документ должен содержать все слова-операнды		

Оператор	Описание	Синтаксис	Тип
<NOT>	Документы, содержащие слова-операнды, исключаются		Унарный
<IN>	Поиск слова, принадлежащего определенному полю документа	Слово <in> поле Слово <in> (поле$_1$, поле$_2$, …)	Бинарный
“фраза”	Поиск точного совпадения фразы	“Слово [Слово […]]”	Модификатор

На поле оператора <IN> при разборе логического запроса не накладывается никаких ограничений, т.е. его смысл интерпретируется непосредственно поисковой машиной. Поле указывает, в какой именно структурной части документа следует искать слово. Например, могут использоваться следующие поля: TTL (заголовок), ABST (резюме, аннотация), REF (ссылки на документы) и т.д.

Кроме обычного набора операторов AND, OR, NOT, большинство систем позволяет использовать операторы NEAR, CTX, SENT, PAR, обеспечивающие контекстный поиск, т.е. уточнение запроса требованием взаимного расположения терминов в документе.

Примеры булевых запросов:
+"Stratec Biomedical" +merger +acquisition
холодная NEAR вода

Разбор логического запроса осуществляется в два этапа:
– лексический,
– синтаксический анализ.

Задачей лексического анализа является выделение из цепочки символов запроса единых синтаксических объектов, называемых лексемами. Лексемами являются конкретные операторы из таблицы 6, круглые скобки (определяют порядок выполнения) и кавычки (служат для задания ключевых фраз).

Список лексем как результат работы лексического анализатора далее поступает на вход синтаксическому анализатору, задача которого – установить принадлежность этого списка структурным условиям, определяющим синтаксис логических выражений. Эти условия задаются в виде контекстно-свободной грамматики, терминальными символами которой и являются входные лексемы.

На выходе синтаксического анализатора (в случае, если входной запрос и, соответственно, список лексем удовлетворяют синтаксису ЛВ) получается дерево логических операторов. Листьями в дереве являются цепочки слов из запроса, а остальными вершинами – идентификаторы операторов. Сыновья вершины-оператора – это операнды этого оператора. Таким образом, входной запрос разбивается на несколько подзапросов (листьев дерева), которые далее обрабатываются отдельно друг от друга и, соответственно, в результате анализа будет получено несколько ПОЗов.

Булевский запрос предполагает нахождение документов, которые удовлетворяют предикату, заданному в запросе. Если операндами булевских выражений являются слова либо фразы (exact phrase), как в большинстве случаев и есть, то определить релевантность документа запросу просто. Она либо 100% либо 0%.

Данная стратегия булевского поиска своей сильной стороной имеет высокую точность – пользователь получает *только* то, что запросил. Сильная сторона данной стратегии является ее же недостатком. Пользователь должен априори *знать,* что в индексе находится, иначе либо его запросы будут приносить слишком мало документов (низкая полнота), либо слишком много (низкая точность). С этой точки зрения данную стратегию можно назвать *неустойчивой,* в том смысле, что два примерно одинаковых по смыслу запроса приносят сильно отличающиеся наборы документов.

Существует небольшая модификация предыдущей стратегии – это стратегия для булевского поиска с оператором AND (разрешен только оператор AND). Данная стратегия учитывает число совпавших операндов и в соответствии с этим числом (иногда в зависимости от внутренних характеристик совпавших операндов) начисляет данному ПОД вес – степень релевантности ПОЗ.

Предположим, мы имеем следующий инвертированный индекс по ключевым словам:

K_1: $D_1, D_2, D_3, D_4,$
K_2: $D_1, D_2,$
K_3: $D_1, D_2, D_3,$
K_4: $D_1,$

Тогда для запроса K_1 AND K_2 AND K_3 релевантность будет вычислена следующим образом:

100% $D_1, D_2,$
66% $D_3,$
33% $D_4,$

Данный подход характеризуется б*о*льшей устойчивостью. Эта стратегия может быть описана при помощи *функции соответствия*. *Функция соответствия* определяет меру релевантности ПОД и ПОЗ, является ядром стратегии поиска. В данном случае величина значения данной функции пропорциональна количеству совпавших терминов.

Существует большое число функций соответствия, выбор той или иной функции зависит, прежде всего, от операндов, т.е. от того, что по природе представляют собой K_1, K_2, K_3, K_4. Например, в системе SMART мера релевантности – это косинус угла между векторами, описывающими ПОД и ПОЗ.

3 Поиск по естественно-языковому запросу

Другой распространенный тип запроса – запрос на естественном языке. Оказалось, что использование традиционных схем поиска, основанных на ключевых словах и дескрипторах, приводит к значительному поисковому шуму (пользователь получает тысячи нерелевантных документов), т.е. к совершенно

неэффективному решению задачи.

С учетом указанных особенностей можно говорить о стратегии семантического (смыслового) поиска, основанного на лингвистической базе знаний и автоматическом анализе текста с использованием семантического уровня глубины естественного языка.

Информационной основой ЛБЗ является корпус текстов, из которого, в частности, автоматически извлекается так называемая ВС-грамматика, являющаяся базовым компонентом лексико-грамматического анализа текста. Результатом грамматического анализа текста является набор грамматических отношений (именная, предложно-именная, глагольная группы и проч.), выделенных из данного текста.

Таким образом, именно грамматические отношения выступают в роли дескрипторов при индексировании документов и запросов, а в дополнение к логическим отношениям между ними используются еще и семантические отношения, что в конечном счете значительно повышает точность и полноту информационного поиска.

Интерфейс пользователя позволяет настраивать поиск на уровень релевантности выдаваемых документов и представлять документы как в явном виде, так и в виде реферата.

Естественно-языковой интерфейс – это вид пользовательского интерфейса, который принимает запросы на естественном языке, а также, возможно, использует естественный язык и для вывода информации (реакции системы на запрос пользователя).

Преимущества ЕЯ-интерфейсов:

– минимальная предварительная подготовка пользователя. Естественный язык является наиболее привычным и удобным средством коммуникации, и именно в силу этого, с ростом эффективности естественно-языковых систем, он, безусловно, будет вытеснять другие виды интерфейсов к ИПС, традиционные в данный момент;

– большая семантическая сила ЕЯ позволяет описывать сущности и явления мира с максимальной степенью детализации;

– большая скорость создания произвольного запроса (отсутствует стадия формального задавания запроса). Как правило, пользователь сразу может сформулировать корректное ЕЯ-представление запроса, поскольку такое представление является самым естественным для человека, тогда как построение запроса на формальном языке, даже с помощью вспомогательных средств, таит множество ошибок, которые можно зачастую исправить, только проанализировав результат запроса.

Недостатки ЕЯ-интерфейсов:

– неоднозначность естественного языка приводит к множественности смыслов. Специфика ЕЯ такова, что часто запрос может иметь несколько смыслов, о которых пользователь в момент задания запроса не предполагает. Формальные же языки лишены проблемы неоднозначности. Это свойство естественного языка приводит к усложнению интерфейсов и методов анализа, в противном случае интерфейс получается слишком примитивным для реального использования;

– недостаточная надежность анализаторов запросов может привести к их неправильному пониманию. Современные естественно-языковые интерфейсы далеко не всегда позволяют диагностировать причины неудач понимания. Причины этих неудач могут быть как в лингвистической сфере, так и в концептуальной (интерфейс должен уметь отличать реальную предметную область, которую имеет в виду пользователь, задавая запрос, от той ее части или трансформации, которая представлена в документе);

Примеры ЕЯ-запросов:

– Вопросительное предложение:

Как нужно заполнять налоговую декларацию?

– Команда (побудительное предложение):

Покажите мне информацию о таможенном оформлении автомобиля.

– Описание проблемы:

Сыну почти 4 года. Замужем я не была, но сын оформлен на отца и фамилия у него отца. Вместе не живем. Могу ли я поменять фамилию ребенка на свою?

Простейшей стратегией поиска по ЕЯ-запросу является сведение ЕЯ-запроса к булевскому; как правило, это стратегия с отказами, т.е. тривиальный алгоритм поиска по ЕЯ-запросу сводится к следующему:

– Выделяются значимые слова в запросе (данная задача, в простейшем случае, может быть решена при помощи словаря стоп-слов).

– Выделенные слова перечисляются через AND.

– Поиск осуществляется по стратегии, описанной выше.

Желание снять с пользователей заботу о формулировании запросов на ИПЯ булевого типа привело к распространению *языков типа "like this"*. Информационный запрос в таких ИПС представляется на естественном языке, однако синтаксический и семантический анализ запроса не производится. Запрос разбивается на слова, из которых удаляются запрещенные (стоп-слова) и общие (предлоги, союзы и т.п.) слова, производится нормализация лексики, а затем все слова связываются либо логическим AND, либо OR (при предположении, что термины, записанные через запятую, представляют собой синонимы, связываемые оператором OR, а через пробел – обязательные термины, связываемые оператором AND). Таким образом, запрос «*Программное обеспечение для платформы Unix»* будет преобразован в запрос «*программное обеспечение AND платформа AND Unix»,* который принадлежит ИПЯ булевого типа.

Как правило, более сложные алгоритмы поиска по ЕЯ-запросу работают примерно по такой же схеме. Стратегия поиска определяется функцией соответствия, а функция соответствия, в данном случае, определяется теми лингвистическими структурами, которые лингвистический процессор способен извлекать из данного ЕЯ. Например, если лингвистический процессор выделяет именные группы, то мы можем сравнивать именные группы запроса с именными группами документа и т.д. Это не может не наложить определенный

отпечаток на структуру индекса в целом.

Функция соответствия для поиска, основанного на ЕЯ, является весьма сложной, как правило, она зависит от самого запроса, т.е. вначале определяется тип данного ЕЯ-запроса, выбирается функция и осуществляется поиск. При вычислении релевантности ПОЗ и ПОД часто пользуются эвристикой наименьшего расстояния. Эвристика заключается в том, что ближе расположенные друг к другу лингвистические отношения являются более связанными между собой, т.е. при прочих равных тот ПОД, чьи совпавшие лингвистические отношения находятся ближе, будет более релевантным ПОЗ.

Это часто упрощает анализ, но существенно снижает релевантность возвращаемых результатов.

Основной целью функционирования подсистемы анализа запросов к ИПС на естественном языке является формирование поискового образа запроса, который должен по возможности адекватно отражать его синтаксис и семантику и быть достаточно информативным для нахождения релевантной поставленному запросу информации.

Сформированный ПОЗ может быть использован далее другими модулями ИПС для ряда подзадач:

– извлечение из базы данных и ранжирование по релевантности документов, вероятно содержащих искомую информацию;

– сопоставление логической структуры документа с ПОЗ (процедура унификации);

– выдача релевантных документов пользователю либо самостоятельное формирование ответа на запрос путем извлечения информации из документов и логического вывода новых фактов из уже имеющихся.

Анализ запроса предполагает извлечение из него по возможности всей доступной информации о его структурных единицах на различных уровнях, поэтому подзадача анализа запроса является частным случаем более общей задачи автоматической переработки текста. Соответственно, к ней применимы перечисленные ниже требования к задаче автоматической переработки текста.

При рассмотрении ЕЯ-запроса к ИПС с точки зрения пользователя, не имеющего понятия о том, как устроена система изнутри, о структурной и логической организации базы документов, (а в случае поиска в сети Интернет так и происходит), нельзя налагать какие либо ограничения на тип самого запроса. Пользователь может задать ИПС корректный запрос, но может и допустить в нем лексические или грамматические ошибки, может задать вопрос, а может просто описать проблему, решение которой он хочет получить. Системе, настроенной на понимание только одного типа запроса, будет сложно “выжить” в условиях жесткой конкуренции на рынке ИПС в сети Интернет. Поэтому разрабатываемая система анализа запроса должна обладать определенной гибкостью и по возможности большой *полнотой охвата типов запросов.*

В зависимости от глубины проводимого автоматического анализа запроса различают системы с опорой на знания (проводится, как минимум, синтаксический анализ) и без опоры на знания (проводится лексический или лексико-грамматический анализ). Тип системы обуславливает алгоритмы,

используемые на этапах поиска и извлечения запрашиваемой информации. Кроме того, в зависимости от глубины анализа может различаться структура ПОЗ. В простейшем случае это список ключевых слов с весами или без, но также в него могут входить логические и контекстные операторы, синтаксические и семантические отношения, представленные в виде списка или иной логической структуры.

Обработка естественно-языковых запросов происходит аналогично процедуре, описанной для ИПЯ булевского типа.

Специфика обработки запроса заключается в том, что он не содержит достаточно информации для снятия присутствующих грамматических и семантических неоднозначностей. Ошибка при разборе (понимании) запроса стоит очень дорого, т.е. ошибочно разобранный запрос – это резкое уменьшение релевантности ответов, не зависящее от качества поиска. Для решения подобных проблем производят не один, а несколько вариантов разбора. Следует также учесть, что текст, вводимый пользователем, содержит значительное число опечаток.

Можно выделить следующие основные этапы автоматического анализа запроса на ЕЯ.

– Проверка на присутствие грамматических ошибок.

– Лексический анализ.

– Удаление неинформативных, вводных частей, характерных для устной речи и не несущих смысловой информации.

– Определение фокуса вопроса и типа ответа, к которому задан вопрос.

– Построение шаблонов ответа.

– Выделение грамматических и семантических отношений.

– Выделение значимых слов.

– Построение ПОЗ.

Исходя из этого возникает необходимость разработки подсистемы анализа запроса пользователя на естественном языке к ИПС. Результатом анализа должен являться ПОЗ, формально отражающий его синтаксис и семантику, на основе которого возможно осуществить семантический поиск в информационно-поисковых либо вопросно-ответных системах.

Построение поискового образа документа на основе его автоматического анализа

Реализация модуля «Построение ПОД» основывается на работе лингвистического процессора. Задача лингвистического процессора – преобразование естественно-языкового текста в некоторый набор элементов, являющихся формальным представлением его смысла.

Классическая структура лингвистического процессора содержит три последовательных блока – для морфологического, синтаксического и семантического анализа текста. Кроме того, при подготовке исходных данных для работы лингвистического процессора может использоваться блок преформатирования и лексического анализа.

При решении задачи разработки двуязычной системы полнотекстового поиска правовой информации используются лексический и морфологический

(лексико-грамматический) анализы.

Результаты лексико-грамматического анализа используются для построения списка информативных слов документа. Этот список и является поисковым образом документа в данной реализации системы.

Информативность слова определяется на уровне его лексико-грамматических кодов. Класс информативных слов устанавливался с помощью экспертного анализа и в него вошли: существительные (собственные и нарицательные), прилагательные, наречия, причастия и глаголы. Принадлежность слов к данному классу определяется автоматически на основе заранее заданного списка лексико-грамматических кодов слов. При этом для конкретного естественного языка могут накладываться определенные ограничения. Эти ограничения чаще всего оформляются в виде так называемых стоп-словарей. Для русского и белорусского языков в состав таких словарей входят неинформативные слова, принадлежащие к так называемым служебным частям речи (предлог, союз, междометие, частица), к модальным и вводным словам и выражениям, а также местоимения и числительные.

Все выделенные информативные слова должны подвергнуться процедуре канонизации в соответствии с принадлежностью слова к конкретной части речи. В качестве канонической формы в соответствии с требованиями индексирования и нормами русского и белорусского языков рассматриваются:

1) для имени существительного – именительный падеж, единственное число;

2) для имени прилагательного (краткого и полного) – мужской род, именительный падеж, единственное число;

3) для глагола – неопределенная форма глагола;

4) для наречия – положительная степень;

5) для местоимения – именительный падеж, единственное число.

В соответствии с решаемой задачей всем частям речи были приписаны ЛГК для определения принадлежности слова к конкретной части речи, получения канонической формы и полной словоизменительной парадигмы.

Таким образом, алгоритм построения ПОД для каждого *j*-го текстового документа из ЭБДПИ, $j = \overline{1,l}$, состоит в построении списка $D^{(j)}$ информативных слов:

Шаг 1. Из текстового документа выбрать только информативные слова. Выбор этих слов выполняется по их ЛГК с фильтрацией неинформативных слов по кодам и по стоп-словарю.

Шаг 2. Для каждого *i*-го слова $w_i^{(j)}$ из списка $D^{(j)}$ выполнить процедуру канонизации.

Шаг 3. для всех информативных слов $w_i^{(j)}$ *j*-го документа, $i = \overline{1,n^{(j)}}$, вычислить их абсолютные частоты $\nu_i^{(j)}$.

Шаг 4. вычислить вес $p_i^{(j)}$ каждого слова $w_i^{(j)}$:

$$p_i^{(j)} = \frac{\nu_i^{(j)}}{\nu_{\max}^{(j)}}$$

Предложенный алгоритм положен в основу двух подходов, реализующих построение ПОДа.

Автоматический анализ запроса пользователя и его перевода с целью построения поискового образа запроса

Обработка естественно-языковых запросов происходит аналогично процедуре, описанной в предыдущем разделе. Специфика обработки запроса заключается в том, что он не содержит достаточно информации для снятия присутствующих грамматических (а зачастую и семантических) неоднозначностей. Ошибка при разборе (понимании) запроса стоит очень дорого, т.е. ошибочно разобранный запрос приводит к резкому уменьшению релевантности ответов, не зависящему от качества поиска. Для решения подобных проблем производят не один, а несколько вариантов разбора.

Для обеспечения двуязычности при информационном поиске выполняется автоматический перевод введенного запроса на другой язык. В нашем случае речь идет о русско-белорусской языковой среде. При этом язык запроса определяется автоматически.

Как введенный запрос, так и его перевод подвергаются процедуре лексического и лексико-грамматического анализа с целью выделения информативных слов из запроса. При этом выполняется удаление неинформативных слов, вводных частей, характерных для устной речи и не несущих смысловой информации. Полученный список также подвергается канонизации. Обозначим такой список через Q, $| Q | = m$.

Информационный поиск

Задачей информационного поиска является нахождение документов (фраз, предложений), релевантных запросу пользователя. Рассмотрим поиск с точки зрения *стратегии*, приносящей релевантные документы, а не с точки зрения представления и хранения данных. Все стратегии основаны на сравнении ПОД и ПОЗ. Часто это сравнение не делается напрямую, например, когда ПОЗ сравнивается с кластерами документов (или, более точно, с профилем, представляющим кластер), а не с самими документами. Разница между стратегиями поиска часто может быть определена из различий в информационно-поисковых языках. Природа языка запроса часто диктует природу поисковой стратегии.

Как уже отмечалось выше, ЕЯ-запрос является очень привлекательным видом запроса из-за минимальной необходимости обучения пользователя. С точки же зрения поиска ЕЯ-запрос обладает не менее важным преимуществом перед булевскими запросами – это большая семантическая сила естественного языка, т.е. по сравнению с булевским запросом ЕЯ-запрос содержит в себе намного больше информации.

Стратегия поиска определяется функцией соответствия, а функция

соответствия, в данном случае, определяется теми лингвистическими структурами, которые лингвистический процессор способен извлекать из данного естественного языка. Например, если лингвистический процессор выделяет именные группы, то мы можем сравнивать именные группы запроса с именными группами документа и т.д. Это не может не наложить определенный отпечаток на структуру индекса в целом.

Функция соответствия для поиска, основанного на естественном языке, является весьма сложной, как правило, она зависит от самого запроса, т.е. вначале определяется тип данного ЕЯ-запроса, выбирается функция и осуществляется поиск.

Для рассматриваемой задачи двуязычного поиска в ЭБДПИ документов, релевантных ЕЯ-запросу, введем следующее определение релевантности.

Релевантным ЕЯ-запросу считается каждый η–й документ из БД, для которого

$$D^{(\eta)} \cap Q \neq \varnothing$$

Обозначим

$$D^{(\eta)} \cap Q = R^{(\eta)}, \quad \left|R^{(\eta)}\right| = K^{(\eta)}, \quad \eta = \overline{1,q}$$

Уровень r_η релевантности ЕЯ-запросу каждого η–го текстового документа, $\eta = \overline{1,q}$ определяем по формуле:

$$r_\eta = \sum_s p_s^{(\eta)}, \quad s = \overline{1,k^{(\eta)}}$$

Все полученные релевантные документы ранжируем в соответствии с убыванием уровня релевантности r_η для каждого η–ого документа.

4 Поиск по заданному документу

Весьма привлекательной, но мало распространенной является возможность поиска документов по запросу в виде документа, т.е. поиск документов, релевантных данному. В этом случае запросом является документ. Особенностью данного типа запроса является большое количество информации, присутствующее в запросе.

В целом поиск по документу производится аналогично поиску по ЕЯ-запросу, особенностью является большое количество информации, которое скорее мешает (отвлекает) поиск в сторону. Поэтому очень важно использовать реферат документа, а не сам документ.

В общем виде при решении данной задачи вычисляются ПОД и ПОЗ одной процедурой. Для увеличения эффективности и качества перед вычислением поискового образа документа производится автоматическое реферирование данного документа.

Почти все научные документы (патенты, статьи и т.д.) строятся по следующему шаблону. Сначала описывается проблема, затем проводится ее анализ и решение. В конце приводятся основные результаты и выводы. Исходя из такого шаблона, реферат предлагается формировать в виде следующих

четырех полей:

ЗАДАЧА – проблема, рассматриваемая в данном тексте;

РЕШЕНИЕ – метод решения данной проблемы;

ОСОБЕННОСТИ – особенности данного решения;

ССЫЛКИ – содержит заголовок документа, авторов и т.д.

Модуль автоматического реферирования текста реализует алгоритмы преформатирования и лингвистического анализа текста. При преформатировании осуществляется разбиение документа на несколько предопределенных разделов, а также исключается из дальнейшего анализа служебная информация, приводимая в начале текста. После окончания лингвистического анализа текста имеется возможность получить лексико-грамматический код каждого слова, синтаксическую структуру предложения в виде дерева фразы, а также соответствующие семантические отношения.

Процесс поиска документов, релевантных некоторому исходному документу, заданному в качестве запроса, описывается следующим образом.

Получают рефераты запроса и документов при помощи модуля построения классического реферата. Из общего количества полученных грамматических отношений в запросе и документах отбирают определенный процент с наивысшей оценкой из поля РЕШЕНИЕ, а также добавляют грамматические отношения, попавшие в ЗАДАЧА, причем данный процент зависит от их количества в документе или запросе.

Обработка запроса в виде документа осуществляется так же, как и индексирование документа, т.е. ПОЗ документа – это ни что иное, как ПОД данного документа. Следует заметить, что с целью увеличения отклика на данный запрос-документ ПОЗ может быть представлен шире, т.е. включать в себя больше семантических и грамматических отношений, чем ПОД для этого же документа.

Поиск по заданному документу можно описать следующим алгоритмом. В нашем случае алгоритм предназначен для поиска документов, релевантных некоторому исходному, заданному в качестве запроса.

В предлагаемом подходе ПОД и ПОЗ представляют из себя набор наиболее информативных грамматических отношений (ГО - САО-объекты) документа и запроса, который получают следующим образом:

– получают рефераты запроса и документов при помощи модуля построения классического реферата, описанного выше;

– из общего количества полученных ГО в запросе и документах отбирают определенный процент с наивысшей оценкой из поля РЕШЕНИЕ, а также добавляют ГО, попавшие в ЗАДАЧА. Причем данный процент зависит от их количества в документе или запросе.

Процент p грамматических отношений, отбираемых в поисковый образ, вычисляется по формуле:

$$\begin{cases} p = 100,\ n <= n_1 \\ p = p_m,\ n > n_m \end{cases}$$

$$p = p_i + (n_i - n) / (n_i - n_{i-1}) * (p_{i-1} - p_i),\ i > 0,\ i < m,$$

где n – число ГО в тексте,

m – число строк в таблице,

i – номер строки в таблице, такой что $n_{i-1} < n <= n_i$

Для того, чтобы оценить релевантность ПОДа и ПОЗа, введем следующие понятия. Определяем расстояние между двумя грамматическими отношениями как сумму штрафов всех несовпавших признаков.

Определяем расстояние между ПОДом и ПОЗом следующим образом. Для каждого отношения из ПОЗа находим максимально соответствующее ей из ПОДа, т.е. такое, расстояние от которого до рассматриваемого элемента ПОЗа минимально. Тогда расстояние D между ПОДом и ПОЗом определяется по формуле:

$$D = \frac{d_1 + d_2 + ... + d_n}{n}$$

где n – количество элементов ПОЗа,

Простейший алгоритм вычисления расстояния между ПОДом и ПОЗом может выглядеть следующим образом:

– Выделяются значимые признаки в документе-запросе (данная задача решается при по мощи построения АР данного документа и выделения значимых лингвистических отношений).

– Аналогичным образом значащие признаки выделяются для индексированного документа.

– Вычисляется суммарное расстояние между ПОЗ и ПОД как среднее от минимальных расстояний соответствующих признаков. Если же порядок признаков фиксирован, то алгоритм вычисления расстояния можно свести к более эффективному алгоритму вычисления наибольшей общей подпоследовательности.

В идеале СИП должен основываться на семантическом индексе как запросов, так и документов, что обеспечивает, во-первых, распознавание в полнотекстовых базах данных основных типов знаний (объектов и классов объектов, фактов и причинно-следственных отношений между фактами), во-вторых, точный смысловой поиск релевантной запросу информации, в-третьих, эффективную кластеризацию самих информационных ресурсов.