ИТ в юриспруденции
Лекция 2. Информационно-поисковые системы

Розина Ирина Николаевна
Д.п.н., профессор кафедры ИИиПМ

Информационно-поисковые системы , ИПС

- Упорядоченная совокупность документов (массивов документов) и информационных технологий, предназначенных для хранения и поиска информации - текстов (документов) или данных (фактов)

- Два вида информационного поиска: *документальный* и *фактографический*

История развития ИПС

- I этап. 1955-1965. Проблема информационного взрыва и попытка ее решить
- II этап. Сер. 1960 - сер.1970. Человеко-машинные библиотечные ИПС. Переход к автоматизированному индексированию.
- III этап. 1975-1995. Системы поиска по свободному тексту
- IV этап. Сер. 1990 - н.в. Развитие Интернета. Глобальные ИПС. Естественный язык ИПС

Документальные ИПС

ИПС, в которых реализуется поиск по тематическим запросам в массиве документов или текстов с последующим представлением пользователю подмножества этих документов или их копий

Документ - информационный объект, зафиксированный на некотором материальном носителе (бумага, фото-, кино-пленка, магнитная лента, оптический диск и т.д.) и предназначенный для передачи в пространстве и времени в системе социальных коммуникаций

Фактографические ИПС

ИПС, реализующие хранение, поиск и выдачу непосредственно фактических данных (научных, технических, экономических характеристик и свойств объектов, процессов, явлений, адресов, наименований, количественных данных и т.п.)

- **Отличаются** степенью предварительной интеллектуальной обработки материала (о чем говорится в документе) - специальная форма представления сведений (фактов) об определенном объекте или классе объектов
- **Способ представления сведений** - регистрация при вводе или извлечение из документов (текстов)

Пример запроса

Какова скорость света?

- Документальная ИПС - статьи и книги, в которых говорится о скорости света
- Фактографическая ИПС - скорость света = 300 000 км/с

Функции ИПС

- Формирование информационной БД (индексированные или неиндексированные тексты)
- Поиск информации по запросу пользователя (поисковому предписанию), т.е. сравнение смыслового содержания запроса со смысловым содержанием хранящихся в БД документов и выделение тех из них, содержание которых соответствует поисковому предписанию
- Представление результата поиска – процесс, выражаемый фразой “*все или ничего*”
- Корректировка, уточнение поискового предписания запроса, выполняемые в случае неудовлетворенности пользователя полученными результатами поиска

Подсистемы обеспечения ИПС

- Информационное (документы, запросы, метаданные, средства их описания)
- Лингвистическое (ИПЯ, правила, критерии)
- Программное (алгоритмы и программные компоненты)
- Техническое (технические средства хранения, поиска, передачи)
- Технологическое (порядок выполнения бизнес-процессов)
- Кадровое (обслуживающий персонал)

Программные компоненты ИПС

- Spider (*паук*) – браузероподобная программа, считывающая HTML-код веб-страниц, имеющих URL
- Crawler (*сборщик или путешествующий паук*) – порождаемый Spider-ом процесс, который углубляет поиск, перемещаясь по всем локальным ссылкам. Скачивает страницы, анализирует их для нахождения перекрестных ссылок, изменений на страницах, определения дальнейшего пути и пр.
- Indexer (*индексатор*) – анализирует веб-страницы, скачанные пауками, определят их тематическую принадлежность, актуальность, популярность у пользователей. По окончании анализа индексирует ресурсы
- Database (*база данных, БД*) – хранилище скачанных и обработанных индексатором страниц
- Gateway (*шлюз*) или Search Engine / Results engine (*собственно поисковая машина*) – принимает запросы пользователей, анализирует их и извлекает результаты поиска из БД

Проблемы использования ИПС Объем и доступность информации

- БД ИПС растет как *логарифмическая функция* – количество веб-страниц возрастает как *степенная функция*
- Не все ресурсы доступны (правила индексации, установленный запрет на индексирование, динамические веб-сайты, мультимедийное наполнение и пр.)
- Общий объем *лежащих на поверхности ресурсов* (surface web) – около 19Тбайт
- *Недоступный, невидимый объем* (deep web, invisible web) – около 7500Тбайт

соотношение 1:400

- По данным компании BrightPlanet 2000 г. – невидимый объем в 500 раз больше

Наполнение Интернета Оформление контента

- Объединение Интернета с другими средствами массовой информации (СМИ) – печать, телевидение, радио и телефон
- Интеграция интернет-технологий – электронная, мобильная коммерция, интернет-трейдинг, э-газеты и э-журналы с печатным аналогом и без него, веб-сайты телеканалов и радиостанций, теле- и радиопередач, IP-телефония и пр.
- Широкий тематический диапазон содержания и оформления дизайна персональных сайтов – различные организации, фирмы и пользователи-энтузиасты

Информационно-поисковые системы (ИПС)

- Автоматические системы *индексирования* информации (— , «сырая» информация)
- Предметно-ориентированные системы с *каталогами* (— , «дистилированная» информация)
- *Гибридные* системы поиска (— , «сырая» информация)
- *Метапоисковые* системы (— , «сырая» информация)
- *Онлайновые* справочники или специализированные системы, БД (||, «дистилированная» информация)

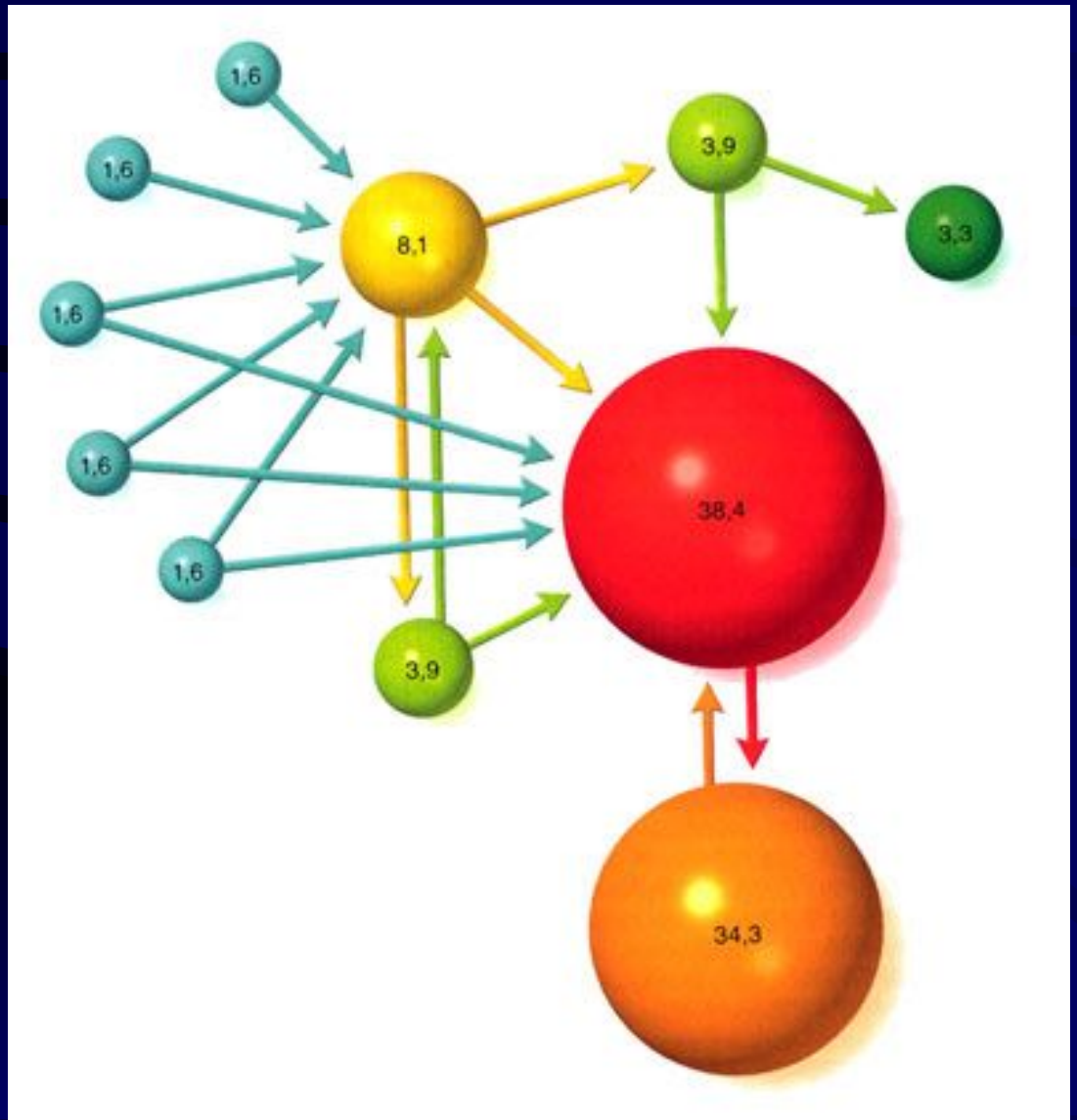
Автоматические системы индексирования информации

Автоматические индексы

- Включают программу-робот (*пауки, индексаторы, агенты поисковые*), БД (*словоуказатели / индексы по текстам документов*) и пользовательский поисковый интерфейс
- Появились в 1994 г. (первые работы в 1960-х гг.)
- Поиск по сочетанию ключевых слов (4-5 ключевых слов)
- Индексация всего содержания страниц (текст, иллюстрации, аудио и видео файлы)

Google, AltaVista, Lycos, Hot Bot, GoTo.com, Excite, InfoSeek, Northernlight, Topping, Anopm, Tela

Алгоритм PageRank приоритизации результатов поиска Google



Тематические (предметные) каталоги / рубрикаторы (Subject Guides, Web directory)

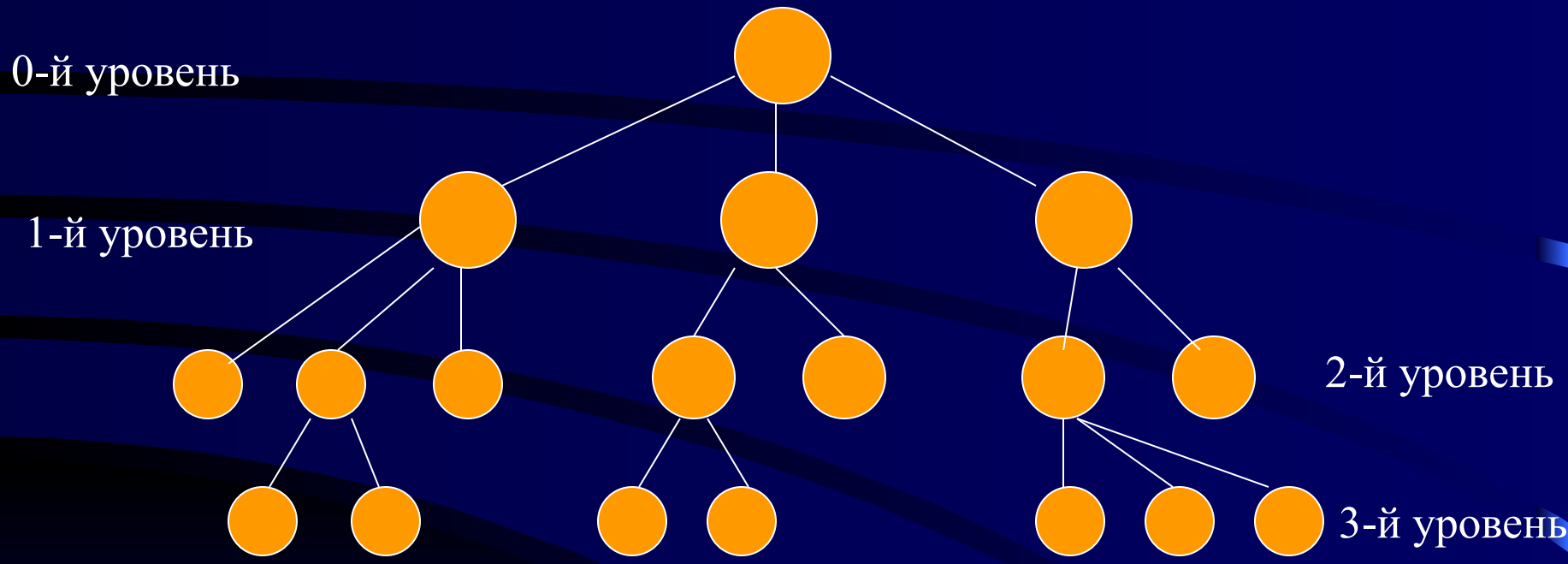
Осмысленно исследуют, описывают, каталогизируют и группируют содержимое WWW-серверов и других ресурсов Интернета при помощи человека-оператора

- Появились около 1995 г. Создание и поддержка требует огромных затрат
- Принцип построения: *от общего – к частному*, иерархический каталог («дерево знаний»): *верхний уровень* общих категорий - бизнес, образование, наука, искусство, путешествия; *нижний* - ссылки на веб-страницы и сервера с кратким описанием содержимого
- Поиск по ключевым словам в кратких описаниях БД
- Объединение с *рейтингами*, учитывающими популярность регистрируемого веб-сайта при помощи *счетчиков посещений*

Yahoo!, Infomine, Virtual Library, Argus Clearinghouse, Galaxy, Look Smart, Net Guide, Snap!, Magellan

Russia on the net, Rambler, Яndex, List, АУ, Refer, Stars, Search, Data, Ulitka

Иерархическая система классификации



Десятичная классификация Дьюи (DDC, Dewey Decimal Classification)

Универсальная десятичная классификация (УДК, UDC, Universal Decimal Classification)

Библиотечно-библиографическая система (ББК)

Специализированные базы данных и поисковые инструменты *Site-specific search engines*

Тематически специализированные объемные БД на WWW, поиск по которым не поддерживается автоматическими индексами

- Базы имен и адресов, библиографические базы данных, цитаты, газетные статьи, словари энциклопедии, информация для трудоустройства и подбора кадров
- Географические карты, информация в области культуры, прогноза погоды, тексты песен, видео продукция, здоровье, бизнес-партнеры

Infomine (БД по различным видам искусства); *InformationPlease* (полнотекстовая энциклопедическая информация); *Университетская информационная система "Россия"* (официальные документы в области экономики, социологии, политологии, международных отношений, *Wayback Machine* (архив веб-сайтов, начиная с 1996 года); *ИНТЕГРУМ-ТЕХНО* (открытые источники информации СМИ, аналитические исследования и обзоры, адресно-справочные и правовые базы данных, информация РОСПАТЕНТа, ГОСКОМСТАТа); *ИНИОН* – Институт научной информации по общественным наукам РАН (аннотированные описания книг и статей из журналов и сборников на 140 языках мира библиотеки ИНИОН).

Интегрированные ИПС

Объединение разных типов ИПС на интегрированной основе

- *Excite, InfoSeek, HotBot, AltaVista,* включают два типа ИПС
- Тематические каталоги *Yahoo!, LookSmart* - поиск в Email directories адреса человека по его имени
- *Google* – поиск изображений, по группам новостей
- *Порталы Yahoo!, Snap! Ампус, Rambler, Yandex, Lycos* - размещение ссылок общего информационно-развлекательного характера – новости, прогноз погоды, фондовые сводки и спортивные результаты и другие ресурсы (энциклопедии, словари, справочники) и сервисы (бесплатные почтовые ящики, веб-страницы, подписка на форумы и пр.)

Метапоисковые ИПС
Metasearch engines, metaengines

Поддержка *метапоиска* - процедуры переадресации заданного условия поиска в другие ИПС

- Запрашивается не более 5–15% БД каждой из используемых ИПС
- Представление результатов поиска на одной странице

All-in-One, LocalFind.com, Search, Dogpile, Baldey, BigHub, SawySearch

www.google.com

- 1998 г. Ларри Пейдж (Larry Page) и Сергей Брином (Sergey Brin), Стэнфордский университет
- Объем индексного файла – 8 млрд. веб-страниц и статей телеконференций
- За сутки индексируется 5 млн. новых и обновленных страниц
- Актуализация каждые 28 дней
- Индексирует документы не только в виде html-файлов, но и в форматах PDF, RTF, DOC, PPT, XLS
- *Google* обозначает 10 в сотой степени

www.google.com.ru

- Русскоязычный интерфейс, а также на 80 языках
- Поиск иллюстраций (425 тыс.)
- Географические карты (GPS)

Расширенный поиск

- *Операторы:* + и –
- *Фразы:* допускаются « »
- *Шаблон:* * не используется
- *Регистр:* учитывается

Яндекс www.yandex.ru

- 1997 г., фирма CompTek
- Объем индексного файла – около 200 млн. документов (около 1,5 млн российских и зарубежных русскоязычных серверов)
- Актуализация каждую неделю
- Индексирует документы не только в виде html-файлов, но и в форматах PDF, RTF
- Морфологический анализ обрабатываемых терминов поискового запроса
- Термин И. Сигаловича, "языковой индекс", "**y**et **an**other in**dex**"
- До 2000 г. - каталог list.mail.ru

www.ya.ru

- Поиск в новостях, собственном каталоге, перечне товаров из электронных магазинов

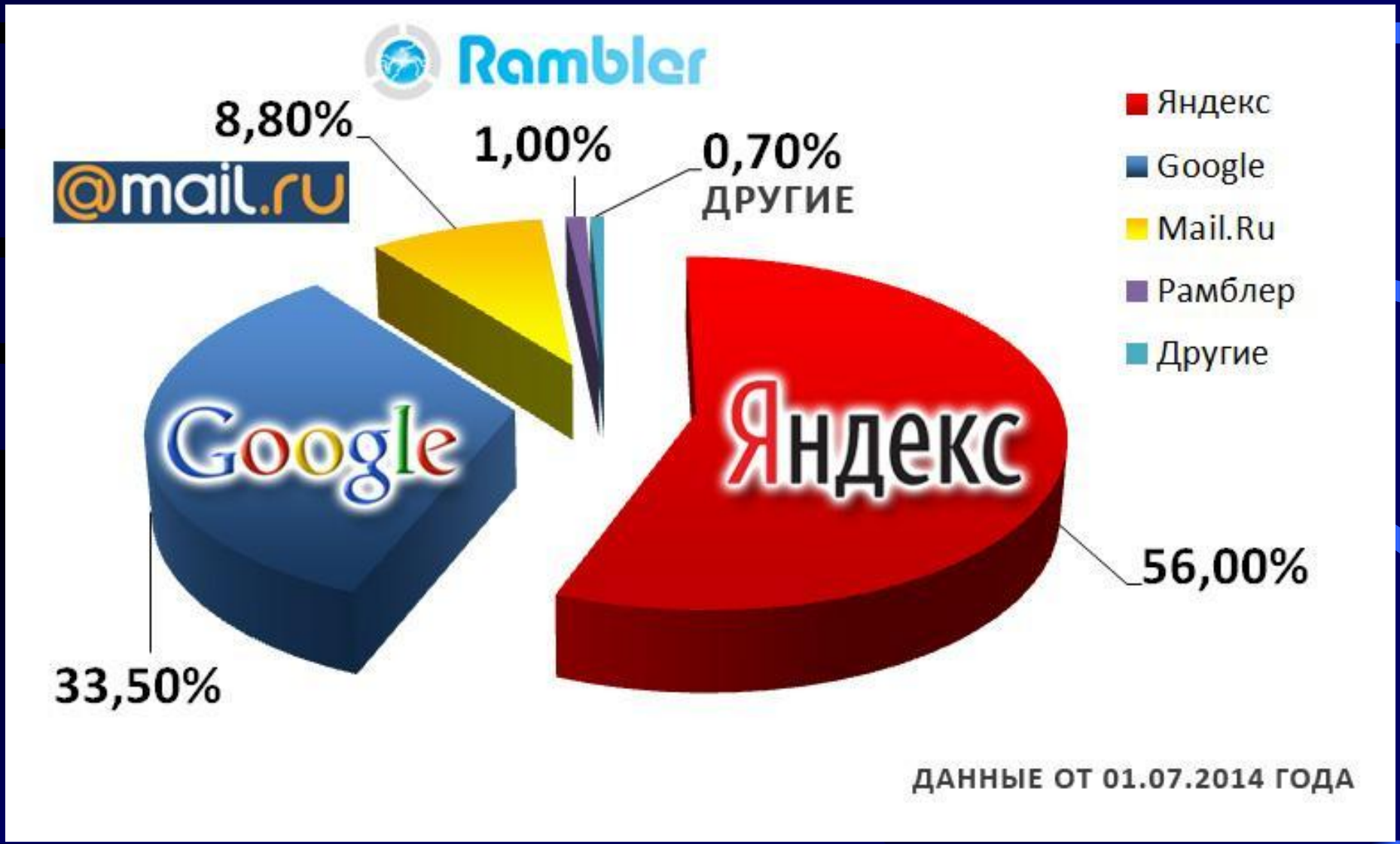
Расширенный поиск

- *Операторы:* + , –, &, &&, |, `, ``, /x (/2 означает расстояние в два слова)
- *Фразы:* допускаются в « »
- *Шаблон:* ! (восклицательный знак перед словом означает точный поиск, без перебора грамматических форм)
- *Регистр:* учитывается

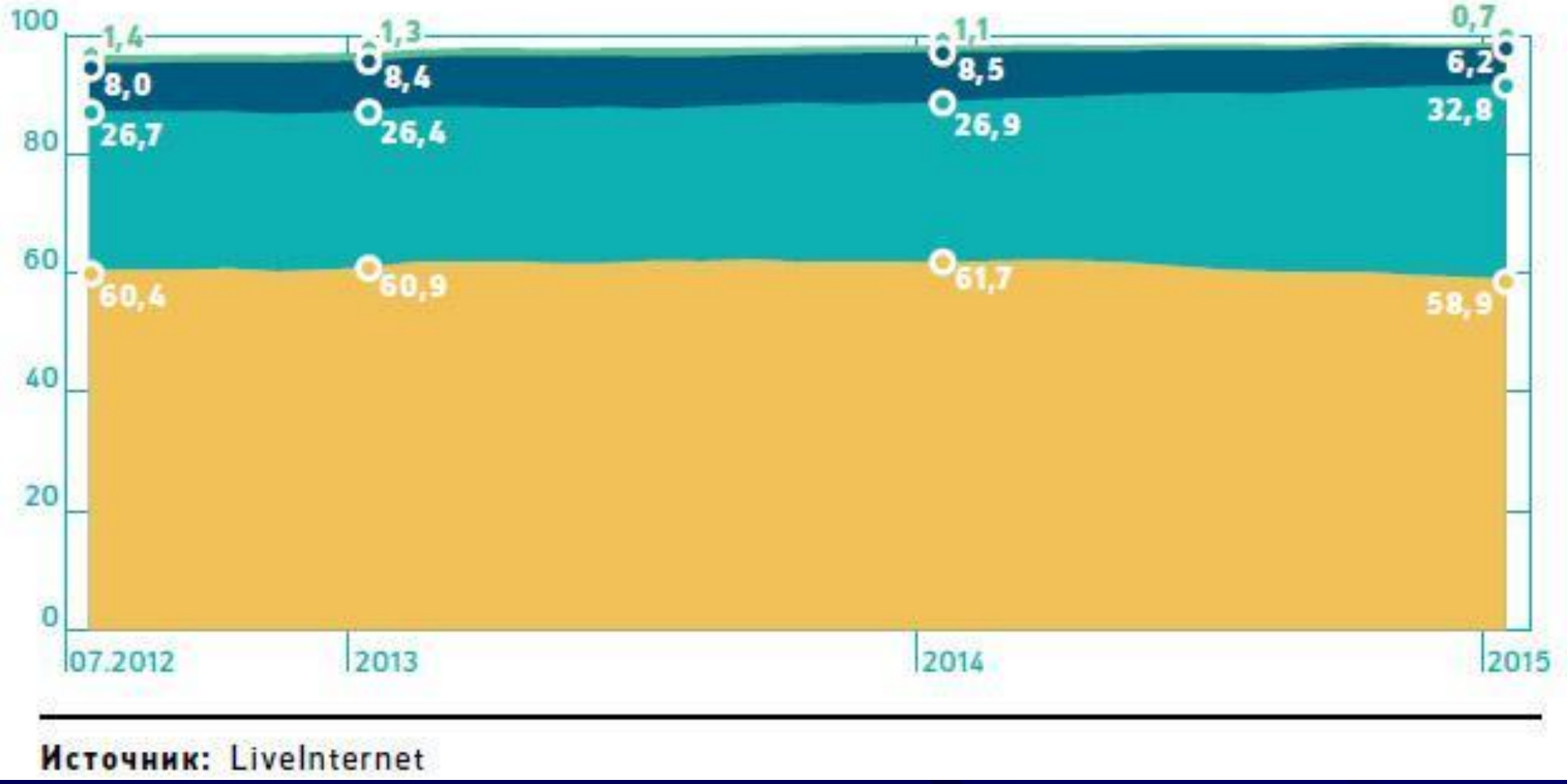
Рамблер www.rambler.ru

- 1996 г., компания Stack Ltd.
- 2000 г. - модернизирован из каталога в портал
- Объем индексного файла – около 120 млн. страниц
- Ежедневно индексирует 6,9 млн. страниц

Популярность ИПС у российских пользователей



Доли поисковиков в трафике российского сегмента интернета, %



Источник: LiveInternet

Оценка функциональной эффективности ИПС

- **Релевантность** (*relevancy*) – критерий отбора информации по тому, насколько *полно* и *точно* тот или иной документ отвечает условиям, указанным в запросе пользователя

$$P = (N1:N)\times 100\%,$$

где N1 – число документов, соответствующих запросу, N – число документов, полученное по запросу, N = N1+N2, где N2 – число документов, не соответствующих запросу

- **Пертинентность** (*pertinency*, *pertinent – уместный*, *подходящий*) – полезность – по отношению к информации, степень соответствия сообщения действительной информационной потребности (прагматическое отношение, субъективная оценка полезности документа пользователем)

Полнота (П) и точность (Т) поиска

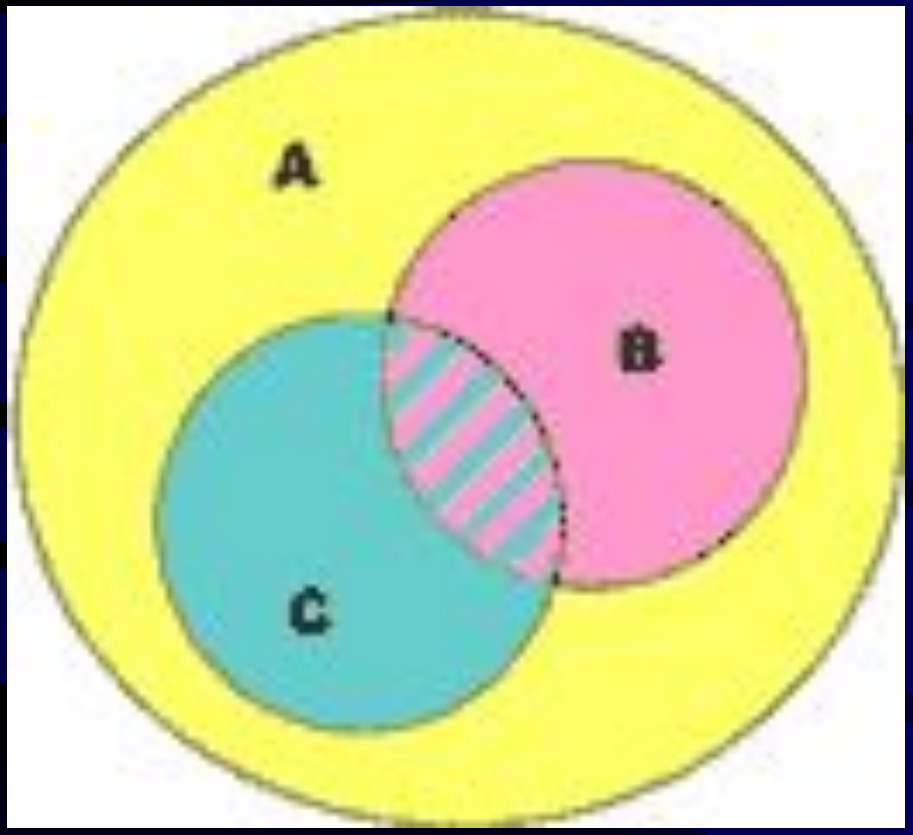
П = (a / (a + c)) * 100 %

Отношение количества выданных релевантных документов к общему числу релевантных документов БД (50-60%)

Т = (a / (a + b)) * 100%

Отношение количества выданных релевантных документов к общему числу документов в выдаче (40-50 %)

Документы	Релевантные	Нерелевантные
Выданные	a	b
Невыданные	c	d



A — множество документов, содержащих ключевое слово.
B — множество релевантных (удовлетворяющих запросу) документов.
(A — B) — множество нерелевантных (шумовых) документов.
C — множество документов, найденных поисковой машиной.
Если C принадлежит B, то точность поиска 100%. Если C = B, то полнота поиска 100%

Оценка ИПС в Интернете

1. Интервал времени между вводом адреса ИПС и завершением загрузки ее главной страницы
2. Объем индексных файлов (общее число проиндексированных серверов и отдельных документов)
3. Возможность быстро сориентироваться и найти поле ввода запроса на странице поиска
4. Интервал времени между нажатием кнопки Поиск и завершением загрузки страницы с результатами
5. Интеллектуальность системы ранжирования результатов поиска
6. Описательность ссылки (возможность судить о найденной странице без перехода по содержащейся в результатах ссылке)
7. Степень оперативности обновления базы данных за счет включения сведений о новых материалах и удаления устаревших
8. Возможности для составления расширенного запроса
9. Наличие дополнительных сервисных функций, облегчающих работу пользователя
10. Наличие излишних элементов, таких как реклама, сложное графическое оформление.

Вопросы

1. Что представляют собой документальная и фактографическая ИПС?
2. Какие функции выполняет ИПС? Назовите программные компоненты ИПС и их функции.
3. В чем состоят технологические особенности **автоматических индексов**?
4. В чем отличие **тематических каталогов** от других ИПС?
5. Назовите технологические особенности **специализированных, интегрированных** и **метапоисковых** ИПС.

6 В чем сложности использования ИПС и чем они вызваны по вашему мнению?