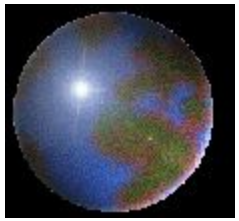
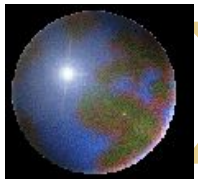


Поиск информации в интернете

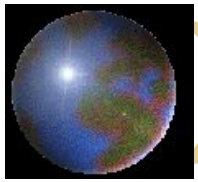
Занятие 1. Вводное



Количество информации в мире растет:

- Калифорнийский университет подсчитал , что в 2002 году в мире произведено

5 000 000 терабайт информации



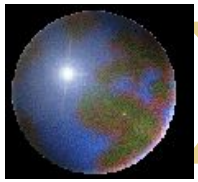
1 терабайт – 1024 Гб

- Для сравнения: объем информации библиотеки Конгресса США, где хранится 19 млн. книг и 56 млн. рукописей –

около 10 терабайт информации

или

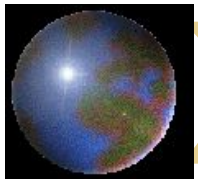
в 500 тысяч раз меньше!



Объем информации в интернете увеличивается в геометрической прогрессии:

- 1998 г. – количество web-сайтов – **около 1 миллиона**
- 2004 г. - web-сайтов – **50 миллионов**, web-страниц – 10 миллиардов

(по данным аналитической компании Cyveillance)

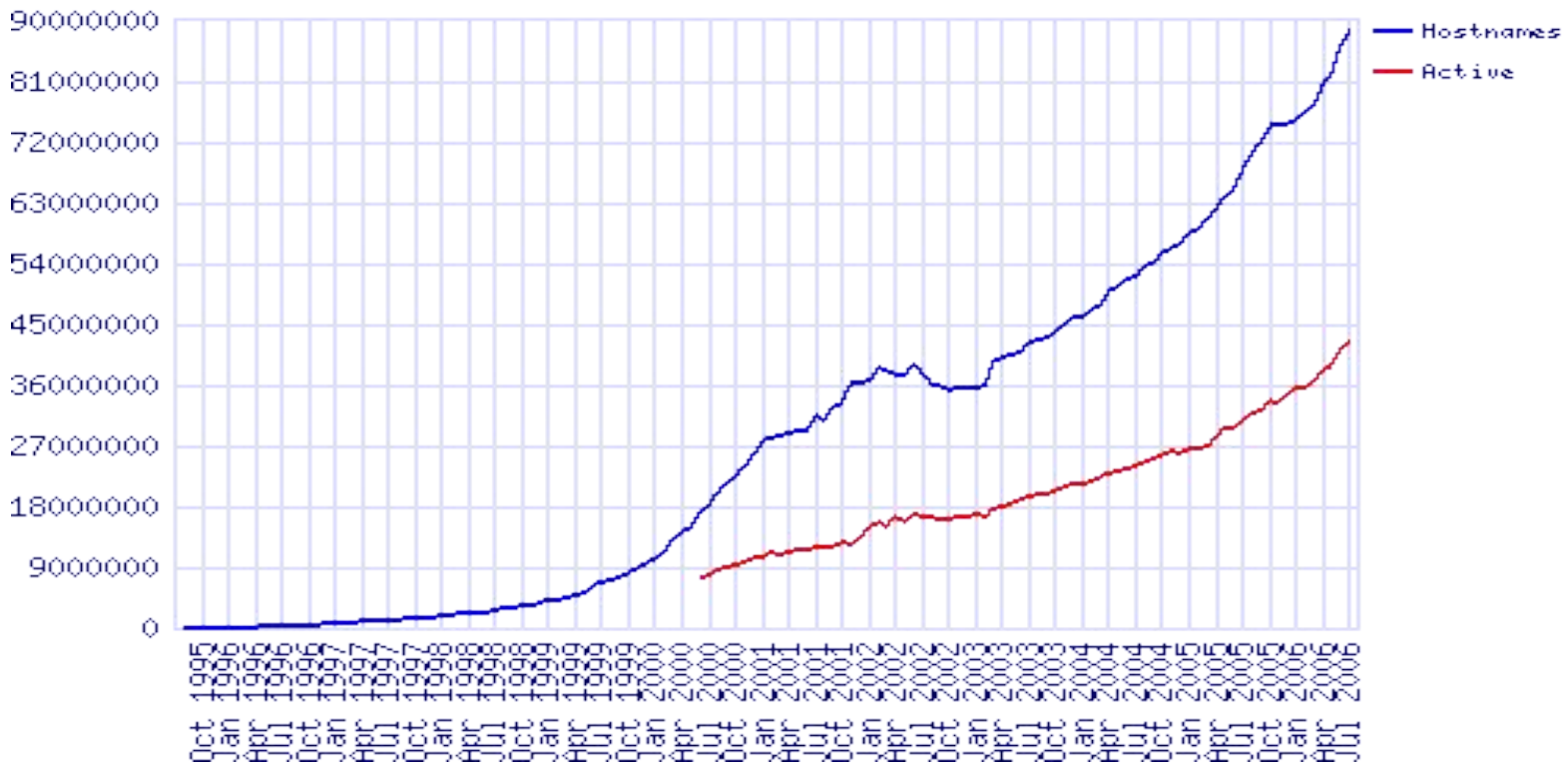


На июль 2006 года:

- По данным аналитической службы Netcraft, в интернете зарегистрировано **88 166 395 сайтов**
- В течение 2006 года количество сайтов увеличивалось примерно на

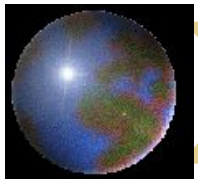
2 миллиона в месяц!

Кривая роста числа сайтов



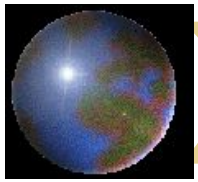
Октябрь 1995 г. – июль 2006 г.

http://news.netcraft.com



Русскоязычный интернет

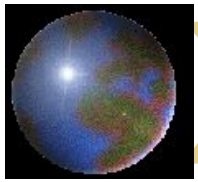
- Аналитики Nigma.Ru в мае 2005 года оценили объем русскоязычного интернета в 1,052 млрд. web-страниц
- А с учетом, т.н. «скрытого Web'а» - не более 1,2-1,3 млрд. страниц
- В то же время специалисты Rambler оценивают объем Рунета в 1,4 млрд. web-страниц



Русскоязычный интернет

В поисковой системе Яндекс на июль 2006 года проиндексировано:

- сайтов: **2 832 533**,
- web-страниц: **1 058 914 756**,
- объем проиндексированной информации: **24 778 ГБ**.

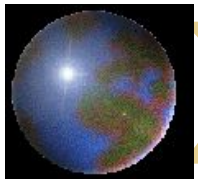


Возникает проблема:

- Переизбыток информации

В США получил распространение «синдром информационной усталости».

По данным исследования Reuters 38% менеджеров «тратят много времени на поиск нужной информации».

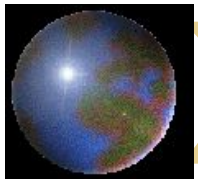


Переизбыток информации

- По данным экспертов Reuters,

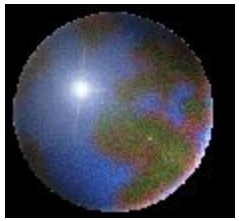
79% журналистов обращаются к интернету в поисках новостей

и лишь 20% находят информацию, которая им необходима!

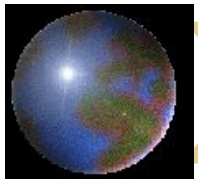


Что необходимо для эффективного поиска информации?

- Представление о структуре интернета.
- Представление о способах и методах поиска информации в интернете.
- Умение сформулировать запрос и выбрать ответ из результатов поиска.



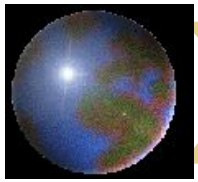
Структура информационного пространства интернета



Благодаря кому в интернете возникает информация?

Как она располагается в интернете?

Как искать, учитывая эти знания?

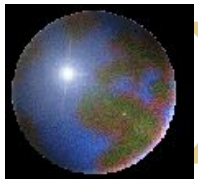


Источники информации

Мы рассмотрим основные источники информации интернета

Особое внимание уделим трем критериям:

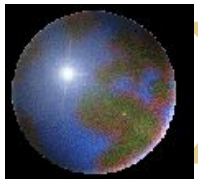
- тематика,
- оперативность,
- достоверность.



Источники информации

#1 Компании и организации (юридические лица), создающие собственные сайты в интернете.

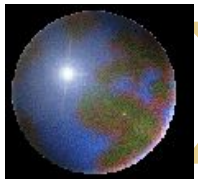
- Тематика, достоверность и оперативность очень широко варьируются



Источники информации

#2 Обычные граждане (физические лица)

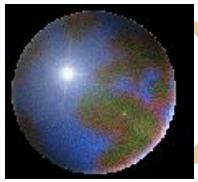
- Чаще всего сайты посвящены увлечению владельца
- Достоверность и оперативность – на совести автора



Источники информации

#2 Они же выступают как **участники форумов, конференций, блогов**

- Тематика – самая разнообразная
- Оперативность – достаточно высокая
- Достоверность – на совести авторов



Источники информации

#3 Журналисты и редакторы сетевых СМИ и информагентств

- Тематика – самая разнообразная
- Оперативность – очень высокая
- Объективность информации зависит от редакции сетевого СМИ (так же, как и у печатных СМИ)

Источники информации

#4 Сотрудники информационных и консалтинговых компаний, создающие специализированные базы данных

- Тематика – самая разнообразная
- Оперативность и объективность – очень высокая

Схема информационных потоков

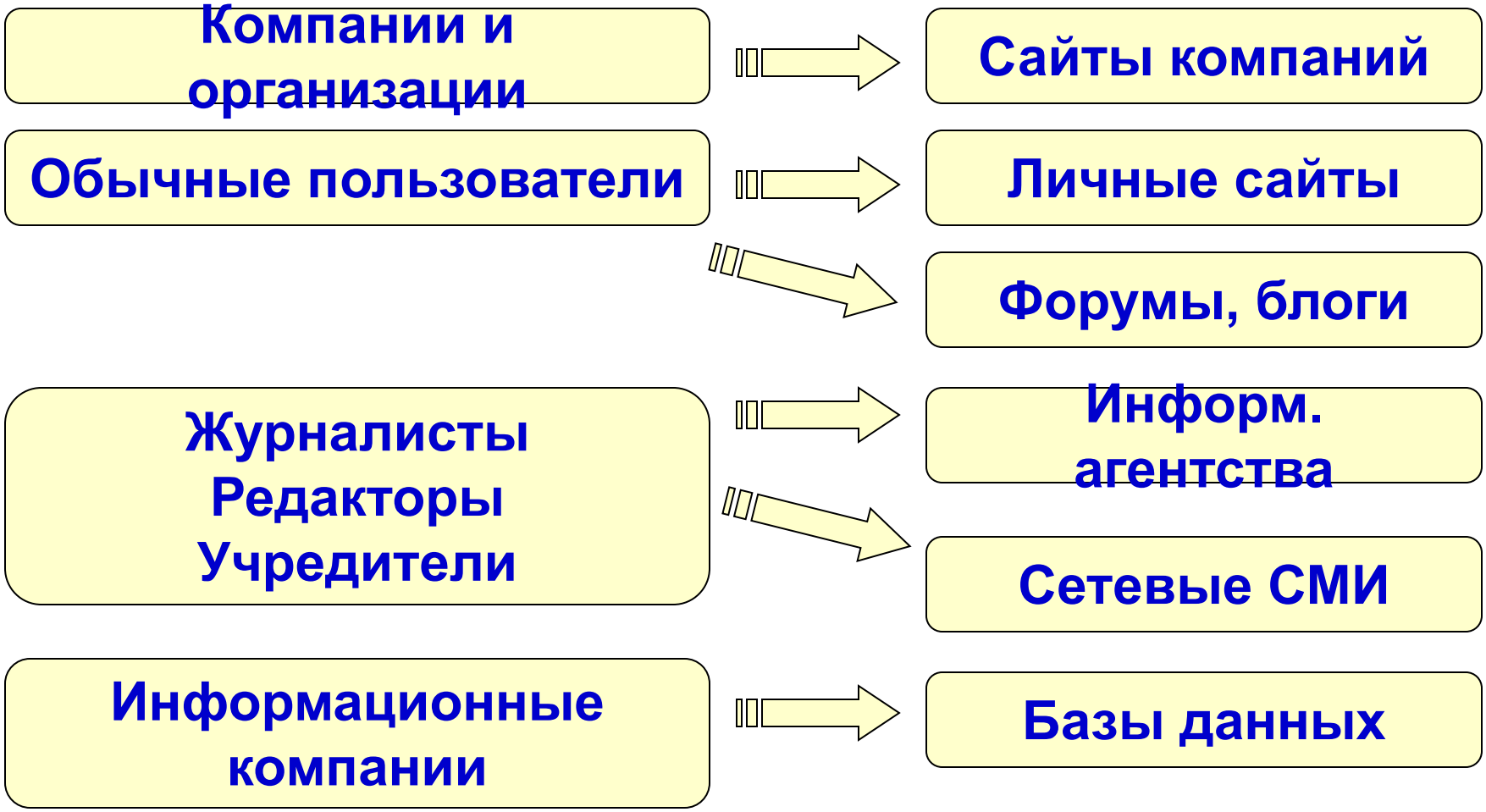
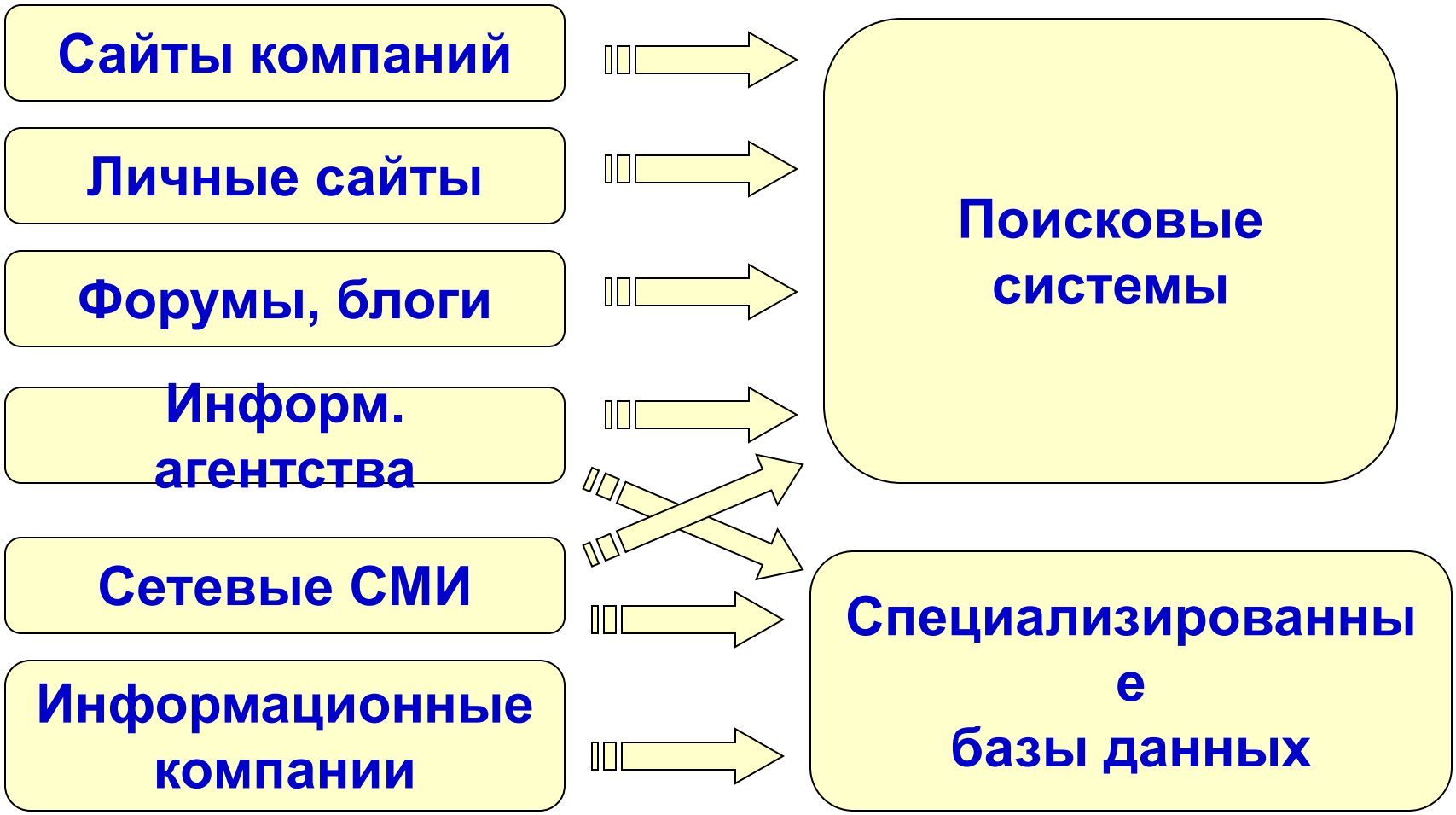
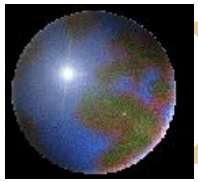


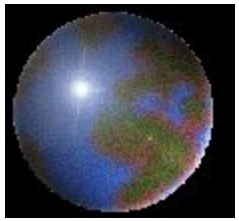
Схема информационных потоков



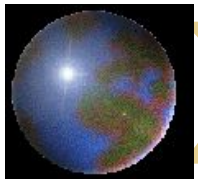


Парадокс интернета:

- Полезной информации становится все больше, а найти что-то необходимое – все сложнее.

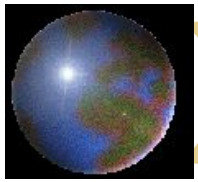


Модель web-пространства



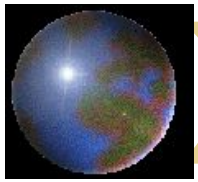
Для эффективного поиска в интернете

- необходимо учитывать архитектуру всего информационного пространства интернета.
- Гиперссылки могут стать основой для построения модели web-пространства.



Модель web-пространства

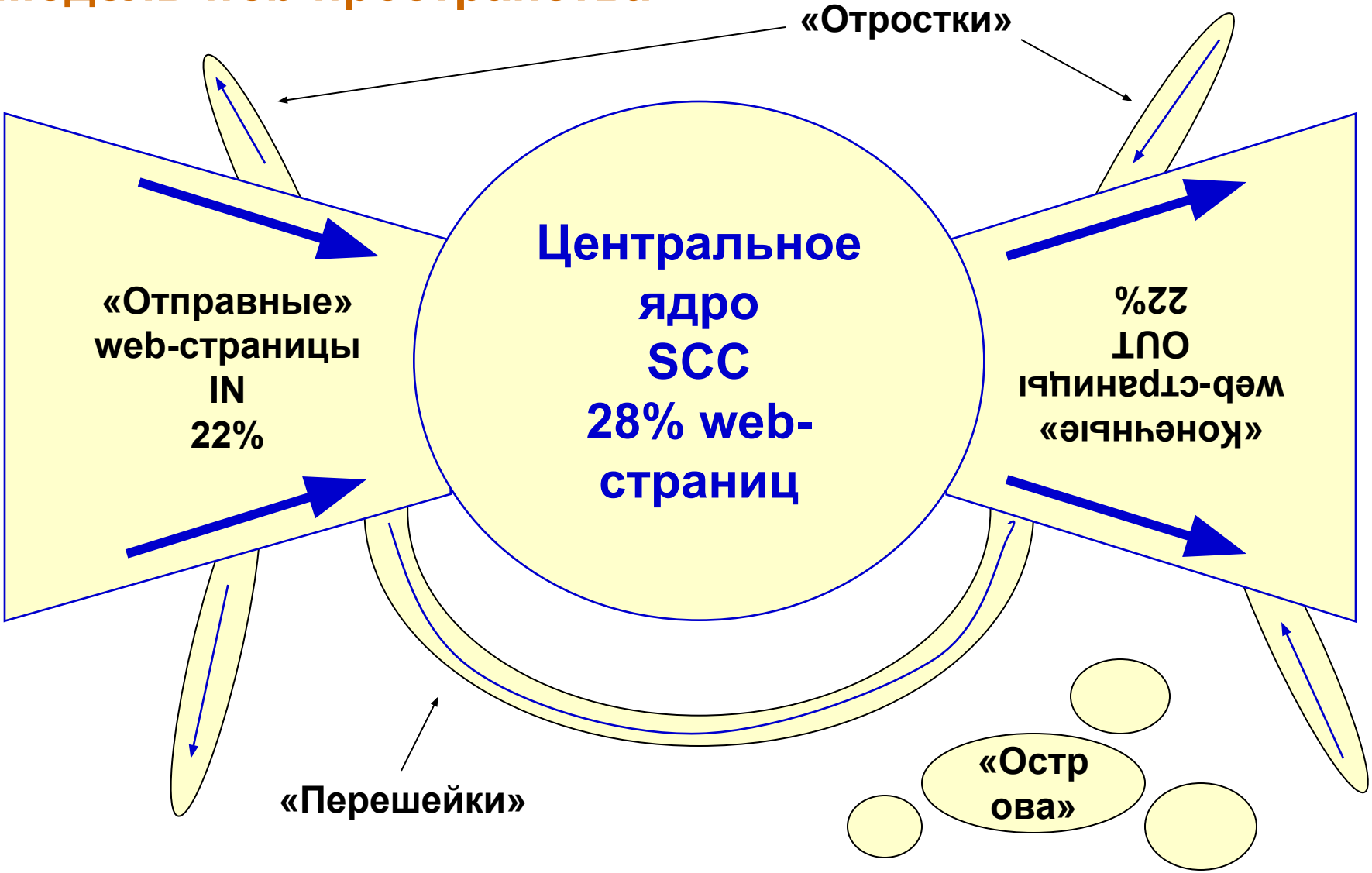
- Впервые создана в 1999 году в Институте поиска и анализа текстов (США).
- Модель опровергла представления об интернете как о едином густом пространстве.

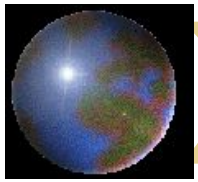


Модель web-пространства

- Проследив с помощью поискового механизма 200 млн. web-страниц и несколько миллиардов ссылок ученые пришли к выводу о неоднородной структуре интернета и создали топологическую модель, близкую к модели Bow Tie (галстук-бабочка)

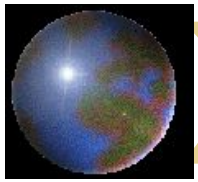
Модель web-пространства





Центральное ядро – 28% web-страниц

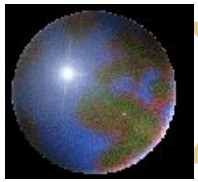
- Компоненты сильной связности (SCC).
- Сюда относятся web-страницы, связанные так тесно, что, следуя по гиперссылкам, из любой из них в конечном счете можно попасть на любую другую.



«Отправные» web-страницы - 22%

- Web-страницы, которые содержат гиперссылки, ведущие в конечном счете к ядру.

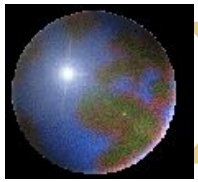
Но! Из ядра по гиперссылкам на них попасть нельзя!



«Конечные» web-страницы – 22%

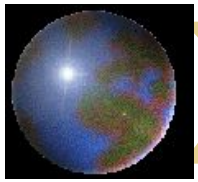
- К этим web-страницам можно прийти по ссылкам из ядра.

Но! Вернуться по гиперссылкам обратно в ядро с этих страниц невозможно!



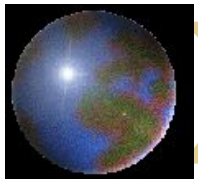
«Отростки» - 22%

- Web-страницы, полностью изолированные от центрального ядра.
- Это либо «отростки», связанные в одностороннем порядке со страницами другой категории.
- Либо «перешейки», соединяющие web-страницы, не входящие в ядро.



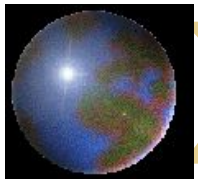
«Острова» - около 10%

- Web-страницы, которые вообще не пересекаются с остальными ресурсами интернета.
- Единственный способ обнаружить эти страницы – знать их адрес.
- Никакие поисковые машины не могут найти «острова», если на них не ведут гиперссылки.



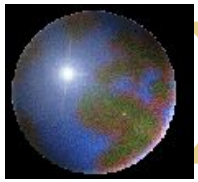
Пропорции модели

- Ученые обнаружили, что пропорции четырех основных категорий web-страниц в течение времени остаются неизменными, несмотря на значительное увеличение общего объема web-ресурсов.



Интернет – это фрактал

- Топология и характеристики модели Bow Tie оказались примерно одинаковыми и для различных подмножеств web-пространства!
- Это позволило сделать вывод о том, что интернет пространство обладает свойствами фрактала.



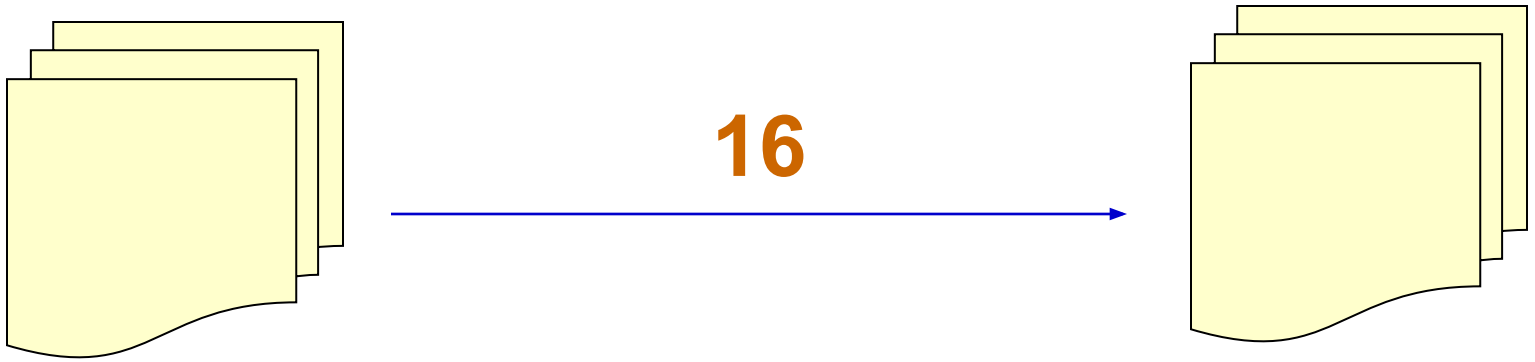
Связь между ресурсами интернет

- Эксперимент выявил сложную картину:

  значительная часть web-пространства отделена от других крупных частей.
- С большой вероятностью случайно выбранные web-страницы окажутся никак не связанными.

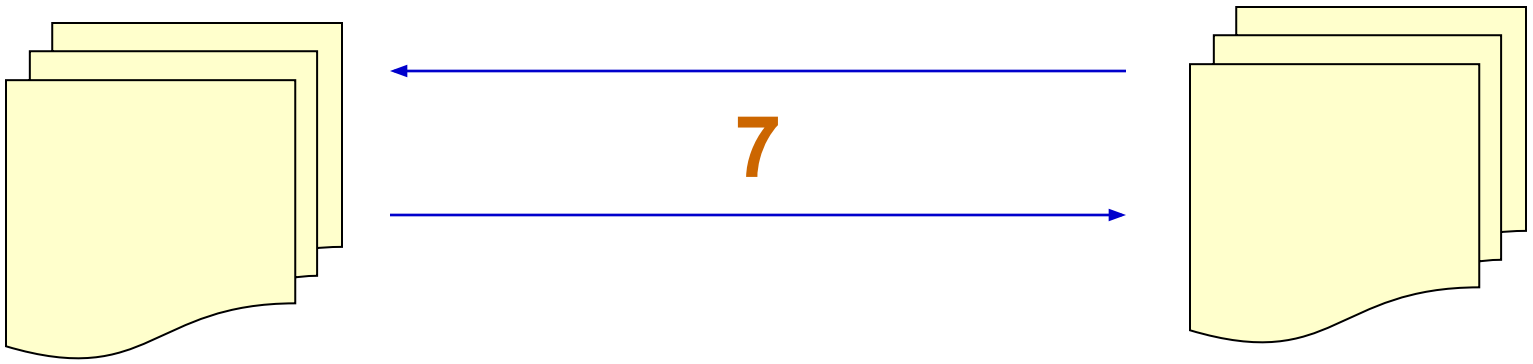
Связь между web-страницами

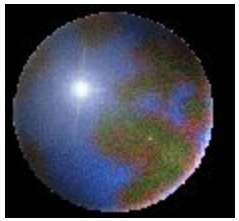
- В случае, если между страницами существует односторонний путь, то среднее количество щелчков для перехода между ними - 16



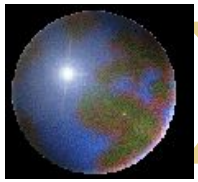
Связь между web-страницами

- Если путь между web-страницами двусторонний, то количество щелчков сократится до 7



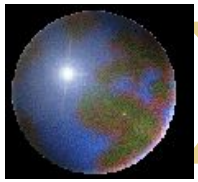


Скрытый Web



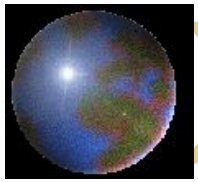
«Острова» - скрытый Web

- Недостаток модели Bow Tie – недооценка размеров «островов», то есть web-страниц, «не видимых» поисковыми системами.
- По оценке компании BrightPlanet в 2000 году число скрытых ресурсов в интернете в сотни раз больше, чем доступных через поисковые системы!



Скрытый Web

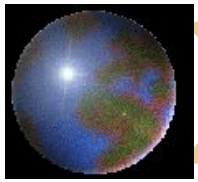
- В 1994 web-ресурсы, недоступные поисковым системам, получили название deep Web или «скрытый Web».
- Другое название этих ресурсов – invisible («невидимый») Web



Скрытый Web

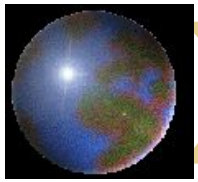
Какие это web-ресурсы?

- Динамически генерируемые страницы
- Информация из баз данных
- Файлы нераспознаваемых форматов
- Системы интерактивного взаимодействия с пользователем
- Платные сайты, защищенные паролем
- и др.



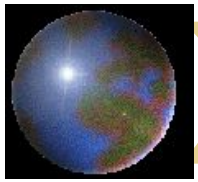
Платные сайты

- Сайты, защищенные паролем и берущие плату за доступ, по некоторым оценкам, составляют всего 10% скрытого Web'а.
- О материалах этих сайтов пользователи ничего не смогут узнать с помощью поисковых систем.



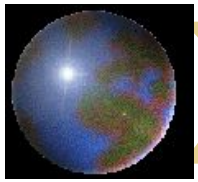
Крупнейшие базы данных

- Одними из самых больших известных ресурсов «скрытого» Web’а являются базы данных служб Dialog и LexisNexis.



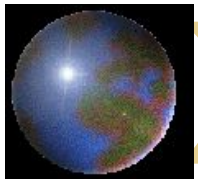
Dialog www.dialog.com

- Создана в 1965 году.
- Dialog содержит 900 баз данных, доступных 700 тыс. пользователей, которые только за один час прочитывают более 17 млн. документов!
- Услугами Dialog пользуются в более чем 100 странах



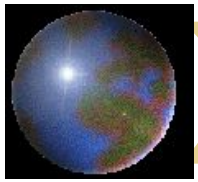
LexisNexis www.lexisnexis.com

- Основана в 1973 году.
- Представляет пользователям юридическую, политическую, коммерческую, новостную и т.п. информацию.
- В первую очередь база данных предназначена для юристов.



LexisNexis www.lexisnexis.com

- Служба охватывает 35 000 источников информации
- 4,6 млрд. документов с глубиной ретроспективы 200 лет.
- В базе содержатся досье более чем на 300 млн. человек!
- Утверждается, что система накапливает только проверенные документы.

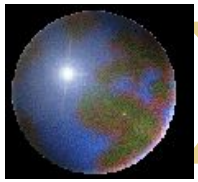


Пример рускоязычной базы данных

- Сайт компании «Кодекс» о российском законодательстве

  www.kodeks.ru

Тысячи документов будут доступны только после входа в систему, поисковые машины не могут проиндексировать содержимое сайта

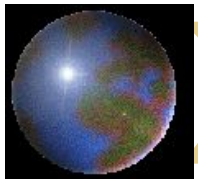


Как искать в «скрытом» Web'е?

- Крупнейший каталог скрытых ресурсов – www.completeplanet.com. Он содержит более 100 тыс. ссылок
- Другие известные каталоги –

  www.bighub.com

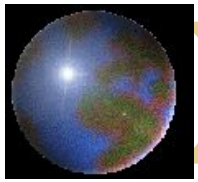
  www.invisible-web.net



Как искать в «скрытом» Web'е?

- Крупнейшая поисковая система для скрытых ресурсов – SurfWax

  www.surfwax.com
- Подавляющее большинство баз данных, доступных в SurfWax относятся к скрытому Web'у.
- Особенность: SurfWax – платная система



Таким образом,

Мы рассмотрели представления исследователей о структуре интернета,

- проанализировали источники информации интернета,
- изучили модель web-пространства,
- описали сущность «скрытого» web’а и возможности поиска скрытых ресурсов