**Понятие гипертекстовой технологии. Гипертекстовая технология WWW, URL, HTML. Структура гипертекста**

Контрольная работа по дисциплине: "Открытые информационные системы"

2005 г.

Гипертекст - текст со вставленными в него словами (командами) разметки, ссылающимися на другие места этого текста, другие документы, картинки и т.д.

Основные компоненты технологии World Wide Web.

К 1989 году гипертекст представлял новую, многообещающую технологию, которая имела относительно большое число реализаций с одной стороны, а с другой стороны делались попытки построить формальные модели гипертекстовых систем, которые носили скорее описательный характер и были навеяны успехом реляционного подхода описания данных.

Идея Т.Бернерс-Ли заключалась в том, чтобы применить гипертекстовую модель к информационным ресурсам, распределенным в сети, и сделать это максимально простым способом. Он заложил три краеугольных камня системы из четырех существующих ныне, разработав:

язык гипертекстовой разметки документов HTML (HyperText Markup Lan-guage);

универсальный способ адресации ресурсов в сети URL (Universal Resource Locator);

протокол обмена гипертекстовой информацией HTTP (HyperText Transfer Protocol).

Позже команда NCSA добавила к этим трем компонентам четвертый:

универсальный интерфейс шлюзов CGI (Common Gateway Interface).

Идея HTML - пример чрезвычайно удачного решения проблемы построения гипертекстовой системы при помощи специального средства управления отображением. На разработку языка гипертекстовой разметки существенное влияние оказали два фактора: исследования в области интерфейсов гипертекстовых систем и желание обеспечить простой и быстрый способ создания гипертекстовой базы данных, распределенной на сети.

В 1989 году активно обсуждалась проблема интерфейса гипертекстовых систем, т.е. способов отображения гипертекстовой информации и навигации в гипертекстовой сети. Значение гипертекстовой технологии сравнивали со значением книгопечатания. Утверждалось, что лист бумаги и компьютерные средства отображения/воспроизведения серьезно отличаются друг от друга, и поэтому форма представления информации тоже должна отличаться. Наиболее эффективной формой организации гипертекста были признаны контекстные гипертекстовые ссылки, а кроме того было признано деление на ссылки, ассоциированные со всем документом в целом и отдельными его частями.

Самым простым способом создания любого документа является его набивка в текстовом редакторе. Опыт создания хорошо размеченных для последующего отображения документов в CERN-е был - трудно найти физика, который не пользовался бы системой TeX или LaTeX. Кроме того к тому времени существовал стандарт языка разметки - Standard Generalised Markup Language (SGML).

Следует также принять во внимание, что согласно своим предложениям Т.Бернерс-Ли предполагал объединить в единую систему имеющиеся информационные ресурсы CERN, и первыми демонстрационными системами должны были стать системы для NeXT и VAX/VMS.

Обычно гипертекстовые системы имеют специальные программные средства построения гипертекстовых связей. Сами гипертекстовые ссылки хранятся в специальных форматах или даже составляют специальные файлы. Такой подход хорош для локальной системы, но не для распределенной на множестве различных компьютерных платформ. В HTML гипертекстовые ссылки встроены в тело документа и хранятся как его часть. Часто в системах применяют специальные форматы хранения данных для повышения эффективности доступа. В WWW документы - это обычные ASCII- файлы, которые можно подготовить в любом текстовом редакторе. Таким образом, проблема создания гипертекстовой базы данных была решена чрезвычайно просто.

В качестве базы для разработки языка гипертекстовой разметки был выбран SGML (Standard Generalised Markup Language). Следуя академическим традициям, Бернерс-Ли описал HTML в терминах SGML (как описывают язык программирования в терминах формы Бекуса-Наура). Естественно, что в HTML были реализованы все разметки, связанные с выделением параграфов, шрифтов, стилей и т. п., т.к. реализация для NeXT подразумевала графический интерфейс. Важным компонентом языка стало описание встроенных и ассоциированных гипертекстовых ссылок, встроенной графики и обеспечение возможности поиска по ключевым словам.

С момента разработки первой версии языка (HTML 1.0) прошло уже пять лет. За это время произошло довольно серьезное развитие языка. Почти вдвое увеличилось число элементов разметки, оформление документов все больше приближается к оформлению качественных печатных изданий, развиваются средства описания не текстовых информационных ресурсов и способы взаимодействия с прикладным программным обеспечением. Совершенствуется механизм разработки типовых стилей. Фактически, в настоящее время HTML развивается в сторону создания стандартного языка разработки интерфейсов как локальных, так и распределенных систем.

Вторым краеугольным камнем WWW стала универсальная форма адресации информационных ресурсов. Universal Resource Identification (URI) представляет собой довольно стройную систему, учитывающую опыт адресации и идентификации e-mail, Gopher, WAIS, telnet, ftp и т. п.

E-mail (Electronic mail) - электронная почта (простонародн. - электронный аналог обычной почты).

Gopher - это интегратор возможностей Internet. Он в удобной форме позволяет пользоваться всеми услугами, предоставляемыми Internet.

WAIS - диалоговая система с оконным интерфейсом для поиска данных по ключевым словам в контексте.

telnet имеет и свой собственный набор команд, которые управляют собственно этой программой, т.е. сеансом связи, его параметрами, открытием новых, закрытием и т.д.; эти команды подаются из командного режима telnet, в который можно перейти, нажав так называемую escape-последовательность клавиш, которая вам сообщается при достижении удаленной машины.

ftp (File Transfer Protocol) - протокол передачи файлов - протокол, определяющий правила передачи файлов с одного компьютера на другой. ftp - также название программы из прикладного обеспечения. Использует протокол ftp для того, чтобы пересылать файлы.

Но реально из всего, что описано в URI, для организации баз данных в WWW требуется только Universal Resource Locator (URL). Без наличия этой спецификации вся мощь HTML оказалась бы бесполезной. URL используется в гипертекстовых ссылках и обеспечивает доступ к распределенным ресурсам сети. В URL можно адресовать как другие гипертекстовые документы формата HTML, так и ресурсы e-mail, telnet, ftp, Gopher, WAIS, например. Различные интерфейсные программы по разному осуществляют доступ к этим ресурсам. Одни, как например Netscape, сами способны поддерживать взаимодействие по протоколам, отличным от протокола HTTP, базового для WWW, другие, как например Chimera, вызывают для этой цели внешние программы. Однако, даже в первом случае, базовой формой представления отображаемой информации является HTML, а ссылки на другие ресурсы имеют форму URL. Следует отметить, что программы обработки электронной почты в формате MIME также имеют возможность отображать документы, представленные в формате HTML. Для этой цели в MIME зарезервирован тип "text/html".

Третьим в списке стоит протокол обмена данными в World Wide Web - HyperText Transfer Protocol. Данный протокол предназначен для обмена гипертекстовыми документами и учитывает специфику такого обмена. Так в процессе взаимодействия, клиент может получить новый адрес ресурса на сети (relocation), запросить встроенную графику, принять и передать параметры и т. п. Управление в HTTP реализовано в виде ASCII-команд. Реально разработчик гипертекстовой базы данных сталкивается с элементами протокола только при использовании внешних расчетных программ или при доступе к внешним относительно WWW информационным ресурсам, например базам данных.

Последняя составляющая технологии WWW - это уже плод работы группы NCSA - спецификация Common Gateway Interface. CGI была специально разработана для расширения возможностей WWW за счет подключения всевозможного внешнего программного обеспечения. Такой подход логично продолжал принцип публичности и простоты разработки и наращивания возможностей WWW. Если команда CERN предложила простой и быстрый способ разработки баз данных, то NCSA развила этот принцип на разработку программных средств. Надо заметить, что в общедоступной библиотеке CERN были модули, позволяющие программистам подключать свои программы к серверу HTTP, но это требовало использования этой библиотеки. Предложенный и описанный в CGI способ подключения не требовал дополнительных библиотек и буквально ошеломлял своей простотой. Сервер взаимодействовал с программами через стандартные потоки ввода/вывода, что упрощает программирование до предела. При реализации CGI чрезвычайно важное место заняли методы доступа, описанные в HTTP. И хотя реально используются только два из них (GET и POST), опыт развития HTML показывает, что сообщество WWW ждет развития и CGI по мере усложнения задач, в которых будет использоваться WWW-технология.

Архитектура WWW-технологии

WWW построена по хорошо известной схеме "клиент-сервер". Программа-клиент выполняет функции интерфейса пользователя и обеспечивает доступ практически ко всем информационным ресурсам Internet. В этом смысле она выходит за обычные рамки работы клиента только с сервером определенного протокола, как это происходит в telnet, например. Отчасти, довольно широко распространенное мнение, что Mosaic или Netscape, которые безусловно являются WWW-клиентами, это просто графический интерфейс в Internet, является отчасти верным. Однако, как уже было отмечено, базовые компоненты WWW-технологии (HTML и URL) играют при доступе к другим ресурсам Mosaic не последнюю роль, и поэтому мультипротокольные клиенты должны быть отнесены именно к World Wide Web, а не к другим информационным технологиям Internet. Фактически, клиент - это интерпретатор HTML. И как типичный интерпретатор, клиент в зависимости от команд (разметки) выполняет различные функции. В круг этих функций входит не только размещение текста на экране, но обмен информацией с сервером по мере анализа полученного HTML-текста, что наиболее наглядно происходит при отображении встроенных в текст графических образов. При анализе URL-спецификации или по командам сервера клиент запускает дополнительные внешние программы для работы с документами в форматах, отличных от HTML, например GIF, JPEG, MPEG, Postscript и т. п. Вообще говоря для запуска клиентом программ независимо от типа документа была разработана программа Luncher, но в последнее время гораздо большее распространение получил механизм согласования запускаемых программ через MIME-типы. Другую часть программного комплекса WWW составляет сервер протокола HTTP, базы данных документов в формате HTML, управляемые сервером, и программное обеспечение, разработанное в стандарте спецификации CGI. До самого последнего времени (до образования Netscape) реально использовалось два HTTP-сервера: сервер CERN и сервер NCSA. Но в настоящее время число базовых серверов расширилось. Появился очень неплохой сервер для MS-Windows и Apachie-сервер для Unix-платформ. Существуют и другие, но два последних можно выделить из соображений доступности использования. Сервер для Windows - это shareware, но без встроенного самоликвидатора, как в Netscape. Учитывая распространенность персоналок в нашей стране, такое программное обеспечение дает возможность попробовать, что такое WWW. Второй сервер - это ответ на угрозу коммерциализации. Netscape уже не распространяет свой сервер Netsite свободно и прошел слух, что NCSA-сервер также будет распространяться на коммерческой основе. В результате был разработан Apachie, который по словам его авторов будет freeware, и реализует новые дополнения к протоколу HTTP, связанные с защитой от несанкционированного доступа, которые предложены группой по разработке этого протокола и реализуются практически во всех коммерческих серверах.

База данных HTML-документов - это часть файловой системы, которая содержит текстовые файлы в формате HTML и связанные с ними графику и другие ресурсы. Особое внимание хотелось бы обратить на документы, содержащие элементы экранных форм. Эти документы реально обеспечивают доступ к внешнему программному обеспечению.

Прикладное программное обеспечение, работающее с сервером, можно разделить на программы-шлюзы и прочие. Шлюзы - это программы, обеспечивающие взаимодействие сервера с серверами других протоколов, например ftp, или с распределенными на сети серверами Oracle. Прочие программы - это программы, принимающие данные от сервера и выполняющие какие-либо действия: получение текущей даты, реализацию графических ссылок, доступ к локальным базам данных или просто расчеты.

Завершая обсуждение архитектуры World Wide Web хотелось бы еще раз подчеркнуть, что ее компоненты существуют практически для всех типов компьютерных платформ и свободно доступны в сети. Любой, кто имеет доступ в Internet, может создать свой WWW-сервер, или, по крайней мере, посмотреть информацию с других серверов.