ПРЕДИСЛОВИЕ

Мы привыкли к Интернету и обращение к нему для многих успело стать чем-то обыденным. Выйти в Интернет, просмотреть новостную ленту, получить и послать e-mail, заглянуть на форум, отыскать новую информацию по профессиональным интересам, разместить в сети что-то свое – для все большего числа людей эти действия превращаются в каждодневную рутину. Но обыденность Интернета обманчива. До сих пор нет единой точки зрения на то, что он есть такое?

Самая распространенная точка зрения заключается в том, что Интернет - это просто самая большая в мире электронная библиотека текстовой, графической, видео- и аудиоинформации практически по любым вопросам. Мы всегда можем подключиться к Интернету и посредством специальных поисковых систем извлечь из него необходимую нам информацию.

С другой точки зрения, Интернет – это некоторая новая реальность, которая предоставляет людям новые возможности по осуществлению политической, экономической, военной, культурной, научной и других видов деятельности. Президенты и правительства, промышленные и финансовые компании, военные и научные организации, учебные заведения, средства массовой информации и даже отдельные физические лица создают в сети Интернет свои представительства, вступают в определенные взаимоотношения друг с другом.

И уж совсем фантастическая точка зрения на Интернет как на материализовавшуюся ноосферу Вернадского, глобальную интеллектуальную систему, новую геологическую силу, которая в скором времени преобразит Землю до неузнаваемости. Не Интернет существует для людей, а мы в определенном смысле существуем для него и являемся всего лишь орудиями его познавательной деятельности. Эта точка зрения лишь кажется такой фантастической, но если присмотреться к сети Интернет повнимательней, то мы обнаружим, что он достаточно автономен, гибель любой его части не ведет к гибели всей системы, что в Интернете существуют активные центры, что в нем протекают процессы обмена информацией, одним из

следствий которых является усложнение и усовершенствование самой глобальной сети.

В настоящей работе мы будем рассматривать сеть Интернет как некоторое глобальное зеркало, которое распростерлось над реальным физическим миром и в котором тем или иным образом, с теми или иными искажениями отражаются события этого мира. Отдельные страницы всемирной сети – это всего лишь *пиксели* на *поверхности* зеркала, а сайты – небольшие группы пикселей. До сих пор, делая запросы к поисковым системам, мы интересовались содержанием отдельных пикселей, но не пытались получить глобальную картину того, что отражено в *зеркале*. В данном случае применимо выражение, что, взаимодействуя с сетью Интернет, мы *за деревьями не видели леса*. Интернет пока что является для нас источником фактов, а было бы хорошо, если бы он стал источником знаний.

Задача, которую мы перед собой ставим, может быть уточнена следующим образом. Пусть дана некоторая модель M_w, которая представляет реальный мир. Требуется построить модель M_i, представляющую Интернет, и определить, какие отношения между этими моделями имеют познавательную ценность, т.е. позволяют на основании свойств структуры M_i делать выводы о свойствах структуры M_w. Важность решения данной задачи состоит в том, что практически все содержание сети Интернет в полном объеме доступно каждому пользователю и требуется лишь научиться его анализировать. Если в физическом мире для уточнения параметров модели M_w нам зачастую приходится проводить ресурсоемкие исследования, то, изучая модель M_w посредством анализа модели M_i, мы практически не расходуем никаких ресурсов. Понятно, что не всякий элемент структуры M_w дублирован в M_i и доступен такого рода анализу, но даже то, что находит отражение в Интернет, все равно поражает своим объемом.

В настоящее время существует направление исследования Интернет, получившее название *web-mining*. Однако круг задач, которые решают в его рамках, в основном ограничен вопросами эффективного поиска, категоризацией текстов, изучением траекторий, по которым перемещаются пользователи глобальной сети, кликая мышкой по гипертекстовым ссылкам. Задачи

интересные, но чисто утилитарные, так как преследуют цель улучшения существующих подходов, а не выход за их рамки.

В числе вопросов, на которые может дать ответ логический анализ, следующие:

1. Какие *типы данных* используются в модели M_i для представления информации о модели M_w?
2. Как представлено *время* в M_i и как оно соотносится с временем M_w?
3. Что есть *событие* в модели M_i?
4. Что значит *существовать* в M_i?
5. Проблема *истинности* в M_i, и ее отношение к истинности в M_w?
6. Каковы *методы рассуждений* над M_i?
7. Каковы *методы поиска закономерностей* в M_i?
8. Возможно ли построение *баз знаний* над M_i?
9. Как *распространяется информация* в M_i?

Полагаем, что приведенный перечень вопросов не является исчерпывающим. Для ответа на них потребуются усилия многих исследователей, но и результат будет стоить того. В настоящей книге мы коснемся лишь части из них, оставив другие для будущих более детальных и глубоких исследований.

АКСИОМАТИЗАЦИЯ ИНТЕРНЕТ

Что мы будем понимать под сетью Интернет?

На самом низком физическом уровне Интернет представляет из себя просто большое число компьютеров, соединенных между собой посредством электрических проводов, оптоволоконных кабелей, каналов радиосвязи и пр. Особого интереса данная структура для логиков не представляет, так как речь идет всего лишь о способе ее технической реализации *в железе*.

На более высоком уровне Интернет состоит не из компьютеров, а из серверов, основная функция которых заключается в хранении информации и ее передаче по определенным правилам (протоколам) другим серверам. Для логиков определенный интерес может представлять анализ протоколов обмена информацией. Здесь находит применение аппарат многосубъектных эпистемических логик. Могут решаться задачи определения логической корректности протокола. Известно, что многие протоколы (наборы правил) обмена информацией между серверами содержат ошибки, которые при определенных условиях могут приводить к некорректной работе. Знание этих недостатков позволяет злоумышленниками получать несанкционированный доступ к различным информационным системам, имеющим связь с Интернет. Логический анализ и устранение таких недостатков является интересной, но все-таки частной задачей.

На еще более высоком уровне, к которому мы собственно и привыкли, Интернет представляет из себя множество сайтов, состоящих в свою очередь из страниц, на которых может быть размещена текстовая, графическая, видео и аудиоинформация. На страницах имеются ссылки, связывающие их с другими страницами и сайтами, что в конечном счете образует гипертекстовую структуру, получившую официальное название World Wide Web – Всемирная Паутина.

Именно последний уровень представления Интернета и будет нас интересовать.

Что существенно для нашего анализа?

Интернет развивается очень бурно. Постоянно совершенствуются способы представления информации на Интернет-страницах, расширяются старые и возникают новые языки для их кодирования. Проблема представления информации также имеет прямое отношение к логике, но в данной работе нас будет интересовать не она. Мы предполагаем, что информация уже тем или иным образом представлена, и задача, которая стоит перед нами, - научиться эффективно пользоваться этой информацией. Поэтому мы отвлечемся от конкретных решений и их реализаций и постараемся принять более общую точку зрения, которая менее подвержена изменениям, связанным с эволюцией Интернет. Нам важно не увязнуть в сиюминутных деталях, а получить результаты, которые останутся значимы еще долгое время.

Более общая точка зрения заключается в том, что Интернет – это реляционная структура, элементарным типом которой являются цепочки символов. Всякая страница сети Интернет – это просто цепочка символов, подчиняющаяся определенному синтаксису. Если мы хотим создать Интернет-страницу, мы должны всего лишь составить некоторый текст и сохранить его на специальном компьютере, подсоединенном к глобальной сети Непосредственно на странице хранится лишь текстовая информация, а графическая, видео и аудиоинформация представлены специальными ссылками на файлы соответствующего формата. Ссылки – это тоже цепочки символов. Специальные программы – интерпретаторы языков, с помощью которых закодированы Интернет-страницы, знают, как найти по ссылкам нужные файлы и представить пользователю в удобном виде содержащуюся в них информацию.. Как это конкретно делается в каждом отдельном случае, для нас совершенно неважно. Важно лишь, что это делается и всегда будет делаться.

Кроме четырех упомянутых выше видов информации в Интернете широко представлена также алгоритмическая информация. Когда мы набираем текст запроса в поисковой системе и нажимаем на кнопку «*Поиск*», мы задействуем алгоритмическую информацию. Некоторые сайты специализируются именно на ней. Описания алгоритмов, которые

при этом используются, также либо закодированы в самой странице, либо представлены ссылками на соответствующие файлы.

Мы принимаем в качестве базового типа данных сети Интернет цепочки символов - слова в определенном алфавите. Базовые операции с ними нам хорошо знакомы. Все остальные, более сложные, типы данных мы должны будем определить с их помощью.

Логическая модель Интернет

Для того чтобы появились цепочки символов, мы должны зафиксировать начальный алфавит букв Alpha, из которых эти цепочки будут строиться. Чтобы не слишком отрываться от действительности, будем считать, что множество букв Alpha конечно. Одним из примеров такого алфавита является хорошо знакомый набор из 256 ASCII-символов. Над этим алфавитом определим множество слов Word:

Def.1
1. Если $a \in Alpha$, то $a \in Word$;
2. Если $v \in Word$ и $w \in Word$, то $vw \in Word$;
3. Ничто другое словом не является.

Базовым отношением на множестве Word×Word является отношение вхождения Include слова v в слово w, которое определяется очевидным образом:

Def.2 Include $\subset$ Word×Word, удовлетворяющее условию
- $\langle v,w \rangle \in Include \Leftrightarrow \exists x,y \in Word(w=v$ или $w=xv$ или $w=vy$ или $w=xvy)$

Мы могли бы определить и другие известные типы отношений и операций над словами, но не станем этого делать, так как их добавление ничего принципиально нового не дает. Важно лишь иметь ввиду, что любые наши действия в конечном счете всегда сводимы к базовым операциям со словами в некотором фиксированном алфавите Alpha.

Мы знаем, что всякое физическое тело имеет пространственно-временные координаты. Нечто подобное

свойственно и Интернет. В нем также имеются свои *тела* - Интернет-страницы как слова в алфавите Alpha, построенные в соответствии с синтаксисом языка HTML или его модификаций.

Def.3 Body $\subset$ Word

Никаких ограничений на размер данного множества мы не налагаем. Важно лишь то, что мы всегда можем эффективно определить, принадлежит некоторое слово b множеству Body или не принадлежит. Это означает, что множество Body рекурсивно.

Как и у физических тел, у Интернет-страниц есть свои *координаты* в пространстве глобальной сети. В качестве координат для пользователей Интернет выступают построенные по определенным правилам URL-адреса страниц, также являющиеся словами в нашем алфавите.

Def.4 Address $\subset$ Word

На размер этого множества мы также не налагаем никаких ограничений и предполагаем лишь рекурсивность.

Заметим, что далеко не каждому элементу множества Address, соответствует реально существующая страница. Пользователям Интернет знакома «*Ошибка 404. Файл не найден*». Это сообщение как раз и говорит о том, что была совершена неудавшаяся попытка перейти по адресу, которому не соответствует ни одна реально существующая страница. В физическом мире тоже не все места в пространстве заняты телами, встречается и пустота.

Помимо этого каждой странице сопоставлено *время* ее создания или последней модификации. Реализуется оно через систему временных меток, которые также являются словами в алфавите Alpha.

Def.5 Time $\subset$ Word.

Множество Time рекурсивно и на нем задано рекурсивное отношение линейного порядка, которое будем обозначать посредством $<$.

Аналогия между физическими телами и Интернет-страницами может быть продолжена. Как и физические тела, страницы глобальной сети *взаимодействуют* друг с другом. Воздействие происходит через посредство ссылок (адресов), и благодаря этому World Wide Web приобретает гипертекстовую структуру и связность.

Интернет-страница появляется тогда, когда некоторый код страницы (тело) размещается по определенному адресу. Это позволяет дать следующее определение:

Def.6 $Page \subset Address \times Body \times Time \times 2^{Address}$, удовлетворяющее условиям:

- $\langle a,b1,t1,R1\rangle \in Page$ и $\langle a,b2,t2,R2\rangle \in Page \Rightarrow b1=b2,\ t1=t2,\ R1=R2$ – функциональность отношения, т.е. страница однозначно задается ее адресом;
- $\langle a,b,t,R\rangle \in Page \ \&\ r \in R \Rightarrow \langle r,b\rangle \in Include$;
- Page конечно.

Следующая интересующая нас структура – это *сайт*, некоторое конечное множество страниц. Сайты характеризуются тем, что у них есть одна так называемая главная страница, адрес которой считается адресом самого сайта.

Def.7 $Site \subset Page \times 2^{Page}$, удовлетворяющее условиям:

- $\langle p,P1\rangle \in Site$ и $\langle p,P2\rangle \in Site \Rightarrow P1=P2$ - функциональность;
- $\langle p,P\rangle \in Site \Rightarrow p \in P$ – главная страница сайта принадлежит самому сайту;
- $\langle p1,P1\rangle \in Site$ и $\langle p2,P2\rangle \in Site \ \&\ p1 \neq p2 \Rightarrow P1 \cap P2 = \varnothing$ - одна и та же страница не может принадлежать одновременно двум сайтам;
- $\{p \mid \exists x \exists P(\langle x,P\rangle \in Site \& p \in P)\} = Page$ – каждая реально существующая страница принадлежит хотя бы одному из сайтов.

Так как каждая страница идентифицируется по адресу в Интернет, возможно альтернативное определение сайтов:

Def.7' $Site \subset Address \times 2^{Address}$, удовлетворяющее условиям:

- $\langle a,A1\rangle \in Site$ и $\langle a,A2\rangle \in Site \Rightarrow A1=A2$ - функциональность;
- $\langle a,A\rangle \in Site \Rightarrow a \in A$ – главная страница сайта принадлежит самому сайту;

- $\langle a1,A1\rangle \in Site$ и $\langle a2,A2\rangle \in Site$ & $a1 \neq a2 \Rightarrow A1 \cap A2 = \varnothing$ - одна и та же страница не может принадлежать одновременно двум сайтам;
- $\{a| \exists x \exists A(\langle x,A\rangle \in Site \& a \in A)\} = \{a| \exists b \exists t \exists R Page\langle a,b,t,R\rangle \in Page\}$ – каждая реально существующая страница принадлежит хотя бы одному из сайтов.

Именно это определение мы и будем использовать в дальнейшем.

Еще одной структурой Интернет, на которую мы хотим обратить внимание, являются домены. Они позволяют объединять различные сайты в тематические группы, задавая на них древовидный порядок. О принадлежности сайта к тому или иному домену можно судить по его адресу, так как составными частями адреса являются имена доменов. В нашем представлении домен – это пара, состоящая из имени домена и множества сайтов, которые ему принадлежат. Внутреннюю структуру имен доменов мы анализировать не будем.

Def. 8 $Domain \subset Word \times 2^{Address}$, удовлетворяющее условиям
- $\langle n,A1\rangle \in Domain \& \langle n,A2\rangle \in Domain \Rightarrow A1 = A2$ – функциональность;
- $\langle n,A1\rangle \in Domain \& \langle m,A2\rangle \in Domain \Rightarrow A1 \cap A2 = \varnothing$ или $A1 \cap A2 = A1$ – множества сайтов, принадлежащие любым двум доменам либо дизъюнктны, либо одно из них является подмножеством другого;
- $\cup\{a| \exists w \exists d(\langle w,d\rangle \in Domain \& a \in d)\} = \{a| \exists x(\langle a,x\rangle \in Site)\}$ – любой сайт принадлежит хотя бы одному домену.

И наконец последним элементом нашей модели Интернет являются поисковые системы. Не будь их, каждому пользователю был бы доступен лишь ограниченный крохотный фрагмент глобальной сети. Именно создателям поисковых систем мы должны быть благодарны за то, какую роль стала играть сеть Интернет в нашей жизни. В ответ на запрос, сформулированный в специальном языке, поисковая система возвращает некоторое конечное множество адресов Интернет-страниц, удовлетворяющих условиям запроса, с указанием на время, когда они были проиндексированы, т.е. занесены в базу данных поисковой системы. База данных поисковой системы является как бы ее внутренним представлением Интернет. Важной

особенностью это базы данных является то, что она принципиально неполна, так как в глобальной сети постоянно появляются новые страницы, но не все они и не сразу заносятся в базу. Одновременно идет и противоположный процесс – страницы исчезают из всемирной паутины, но упоминание о них все еще хранится в базе.

Для начала нам необходимо определить множество слов-запросов Request, посредством которых пользователь дает поисковой системе задание найти те или иные страницы.

Def.9

1. Если w –слово в алфавите Alpha-{ ∧,#,-,),(}, то w∈Request;
2. w∈Request и v∈Request ⇒ (w∧v) ∈Request, (w#v) ∈Request, (w∧-v) ∈Request;
3. Ничто иное словом-запросом не является.

В качестве образца мы взяли языки запросов таких поисковых систем Интернет как AltaVista, Rambler и Яndex. Интересно обратить внимание на используемый в них язык. В нем присутствуют связки конъюнкции ∧, дизъюнкции # и отрицания -. При этом на использование отрицания налагается ограничение. Его можно использовать лишь вместе с конъюнкцией. Т.е. фактически используется не само отрицание, а в язык вводится третья бинарная связка ∧- со смысловой интерпретацией '… и не …', которая в классической логике выражает то же самое, что и отрицание импликации. Интересной особенностью данного языка является то, что в нем невозможно выразить универсально значимое высказывание, т.е. невозможно сформулировать такой запрос, ответом на который было бы множество ссылок на все проиндексированные в поисковой системе страницы.

Определение поисковой системы будет выглядеть следующим образом:

Def.10 SE ⊂ Request×Address×Time, для которого дополнительно выполняются условия:

- <q,a,t1>∈SE&<q,a,t2>∈SE ⇒ t1=t2 – в базе данных поисковой системы хранится лишь время последней модификации страницы;