

Е. Ю. Хрусталеv (Москва, ЦЭМИ РАН). Математическая модель гипертекста.

Принципы построения информационных моделей и методы систематизации сведений для разных областей знаний различаются из-за специфических особенностей информации каждого конкретного вида. Для систематизации обширного класса сведений может использоваться новый тип информационных моделей — гипертекст или нелинейный текст, совмещающий положительные свойства энциклопедии, монографии и тезауруса. Гипертекст обладает рядом характеристик, свойственных как тексту, так и фонду, и не может быть отождествлен ни с одним из существующих методов систематизации информации.

Обычному (одномерному) тексту, который можно интерпретировать как длинную строку символов, читаемую в одном направлении, противопоставлен многомерный текст. В отдельных точках такого ветвящегося многомерного текста чтение можно продолжать в нескольких различных направлениях в зависимости от информационной потребности. Гипертекст может отличаться от обычного текста порядком следования материала, элементы гипертекста могут размещаться в виде иерархического дерева или сетевой организации, он может иметь несколько уровней краткого изложения и детализации материала, способов его представления и т. д.

Введем следующие обозначения. Пусть M и R — конечные непустые множества: $M = \{m_1, m_2, \dots, m_n\}$; $R = \{R_1, R_2, \dots, R_k\}$.

Условимся элементами множеств M и R обозначать объекты и отношения между ними соответственно. Тогда гипертекст в наиболее общем виде (Γ_o) описывается как совокупность следующих четырех компонентов: $\Gamma_o = (T, I, S, Q)$, где T — тезаурус гипертекста; I — информационная составляющая гипертекста. I включает в себя содержание информационных статей I_i , в которые помещаются сведения о всех $m_i \in M$, т.е. $I = \cup I_i$; S — алфавитный (или хронологический) словарь всех наименований $m_i \in M$; Q — список главных тем гипертекста.

Тезаурус гипертекста состоит из тезаурусных статей, каждая из которых может быть представлена в виде: $t_i = \{m_i, Am_i\}$, где t_i — тезаурусная статья объекта m_i ; Am_i — множество объектов, с которыми m_i связан отношениями из R с указанием типа отношения.

Например, если объект m_a связан с другими объектами из множества M отношениями $R_i, R_j, \dots, R_e$, т.е.

$m_{i1}, m_{i2}, \dots, m_{it}$ связаны с объектом m отношением R_i ,

$m_{j1}, m_{j2}, \dots, m_{js}$ связаны с объектом m отношением R_j ,

.....

$m_{e1}, m_{e2}, \dots, m_{ep}$ связаны с объектом m отношением R_e ,

то $Am_a = \{R_i(m_{i1}, m_{i2}, \dots, m_{it}), R_j(m_{j1}, m_{j2}, \dots, m_{js}), \dots, R_e(m_{e1}, m_{e2}, \dots, m_{ep})\}$.

Здесь существенно отметить, что в Am_i содержатся только ближайшие «родственники» для m_i .

Всю совокупность тезаурусных статей, т.е. тезаурус гипертекста можно изобразить следующим образом:

$$T = \{m_1, m_2, \dots, m_i, \dots, m_n, Am_1, Am_2, \dots, Am_i, \dots, Am_n\}.$$

Тезаурус гипертекста графически можно представить в виде сети, узлы которой содержат текстовые описания объектов, а ребра указывают на существование связи между объектами и позволяют определить тип связи.

Наиболее важным компонентом гипертекста Γ_o является совокупность гипертекстовых статей Γ : $\Gamma = \cup_i \Gamma_i$, $\Gamma \subset \Gamma_o$.

Статья Γ_i , состоящая из описания объекта m_i (информационная статья I_i) и перечня родственных объектов (тезаурусная статья t_i), может быть представлена в виде: $\Gamma_i = \{t_i, I_i\}$.

Выдача информации и ввод новых сведений осуществляются, как правило, Γ_i -порциями.

Для аналитического осмысления и последующей автоматизации процедуры поиска семантических сведений в гипертекстовом массиве необходимо составить математическое описание этой процедуры. Важность этой задачи подтверждается также тем, что проблема поиска возникает при внесении новой информации в гипертекст. В этом случае следует выделить все родственные объекты для того, чтобы установить требуемые тезаурусные связи.

Для описания процедуры поиска могут применяться следующие модели:

- 1) поиск ближайшей по составу и содержанию гипертекстовой статьи:

$$\rho(\Gamma_i, \Gamma_c) \rightarrow \min_{\Gamma_i \subset \Gamma}, \quad (1)$$

где ρ — некоторая характеристика близости состава и содержания гипертекстовых статей; Γ_c — вводимая статья или статья, которую ищут;

- 2) поиск гипертекстовых статей с наиболее желательными свойствами:

$$f(\Gamma_i) \rightarrow \max_{\Gamma-i \subset \Gamma}, \quad (2)$$

где f — некоторые желательные свойства, например, наличие одинакового родового объекта;

- 3) комбинированный случай:

$$f(\Gamma_i) \rightarrow \max_{\Gamma-i \subset \Gamma}, \quad \rho(\Gamma_i, \Gamma_c) \leq \rho_o, \quad (3)$$

где ρ_o — некоторое ограничение.

Соотношения (1)–(3) применяются не только при поиске сведений по запросам абонентов, но и для нахождения ранее сформированных статей в целях установления с ними перекрестных ссылок при поступлении в гипертекст новой информации.

Разработанная математическая модель гипертекста и основных операций манипулирования с ним позволяет формализовать перспективные подходы к теоретическому решению интеллектуальных проблем создания и ведения гипертекстовых массивов семантических данных.

Работа, представленная данным сообщением, подготовлена при поддержке РГНФ, проект № 08–02–00255а.