Сокращение читаемого текста без потери смысла. Семпрессия — как и зачем

Табачук И. С. (tabachuk@2tgroup.com)
ООО «2t Group», г. Москва

Непрерывное увеличение потока информации, требующей своевременного ознакомления и\или реакции, неуклонно возрастает. Современная деятельность — научная, управленческая и многие другие, требуют быстрой классификации и сортировки входящих сведений по важности и актуальности. При этом простого реферирования текста недостаточно, поскольку реферат, даже очень хороший, подразумевает потерю или искажение части информации, что порой бывает категорически недопустимо.

Технология, позволяющая сокращать объем сведений и данных, которые могут быть представлены в виде выборки фрагментов, наиболее важных для понимания текста, требующего прочтения, является весьма актуальной.

Главные требования к данной технологии — технологии смыслового сжатия — могут быть сформулированы следующим образом:

1. Входной документ должен быть доступен в исходном виде без искажений.

2. Фрагменты текста, составляющие основной смысл содержания, должны быть идентичны источнику (не являться их пересказом).

3. Должен обеспечиваться прямой доступ от представленного для чтения фрагмента текста к его контексту.

Можно предположить, что наиболее подходящим решением может быть документ в виде гипертекста, использующего разметку, и который будет изначально представлен в сокращенном объеме, а также отображать только наиболее важные фрагменты, содержащие основной смысл документа. При этом представленные фрагменты будут обеспечивать переход к контексту, в котором они встречаются в исходном тексте, обеспечивая легкий доступ к деталям в случае необходимости.

Поскольку контекст основных фрагментов текста может так же быть представлен в виде ссылок на свой собственный контекст, получается многослойный текст, состоящий из главного первого слоя, содержащего самое важное в тексте (если оно есть), и последующих уточняющих слоев, содержащих детали и дополнения.

Фактически, такое представление можно считать смысловым сжатием без потери смысла (рис. 1).

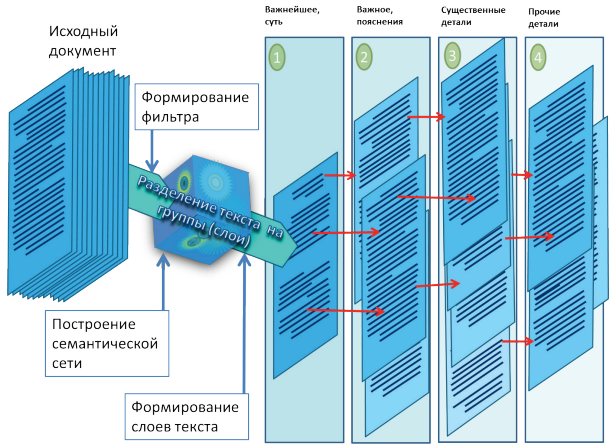

Рис. 1.

Практически формирование многослойных документов возможно на основе технологи построения нейроморфных семантических сетей. Примером построения и использования таких сетей служит хорошо зарекомендовавший себя «ТекстАналист» — программа, созданная компанией «Микросистемы» [ссылка, если необходимо].

На рис. 2 приведен пример фрагмента текста, преобразованного в семантическую сеть. Каждый узел имеет размер, пропорциональный значимости слова в исходном тексте, а также связи, которые отражают его связность с другими узлами сети.

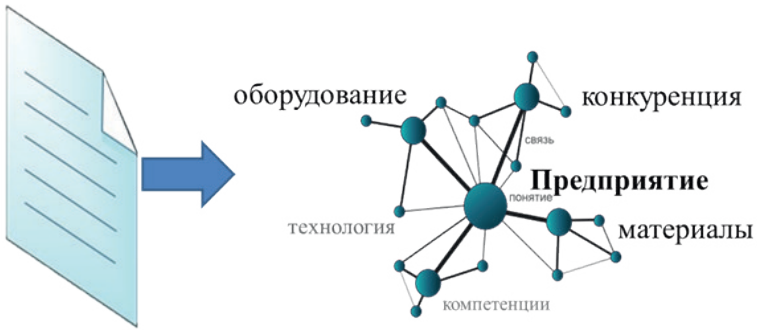

Рис. 2.

То, что формирование пространственного графа формализовано, дает возможность избежать субъективности в определении главных смыслов, содержащихся в тексте. Это особо ценно тем, что на отображение главного не влияют намерения авторов — то, что они сами сказать хотели, а отображается исключительно то, что они сказали фактически.

Численные методы, заложенные в основу технологии построения семантических сетей, позволяют создать настраиваемые фильтры для формирования слоев,

разделяя их по важности для понимания текста. Дополняя фильтрацию по смыслу возможностью поиска фрагментов, которые могут заранее представлять интерес, можно вообще исключить документы, не представляющие интереса в данный момент времени из списка подлежащих прочтению.

Одно из преимуществ такого подхода по сравнению с ручной обработкой текстов заключено в том, что в качестве поисковых моделей, включенных в процесс фильтрации документа и построения его слоев, могут быть использованы не только отдельные слова и/или фразы, но и целые предметные области, которые могут рассматриваться, как единый документ.

Физически это могут быть документы, группы документов, базы данных и прочие хранилища информации или их части и разделы.

Довольно часто формирование запроса, что мы хотели бы найти (в группе документов или отдельном документе) представляет определенные трудности и не позволяет обойтись формальным перечислением слов. К тому же, перестановка слов и даже целых предложений, хоть и не меняет содержание запроса, но может менять его смысл.

Использование главных страниц документов, синтезированных путем смыслового сжатия, могут в свою очередь использоваться в качестве поискового запроса в других предметных областях. Это может дать неожиданные результаты, позволяя выявлять закономерности, которые ранее были скрыты от глаз или просто не замечались.

Почему рассмотрение семантического сжатия приводит к решению задачи поиска информации — вполне логично (Рис.3). Ведь для группы документов, предметной области, рассматриваемой, как единый документ, может быть сформирована своя «главная страница» с заданной степенью детализации (сжатия). Используемая в виде семантической сети, такая структура может служить в качестве запроса для поиска. При этом, что важно, для обеспечения пертинентности, более глубокие слои «сжатого» текста могут использоваться для автоматического уточнения контекста, в котором найдены совпадения результата и запроса.

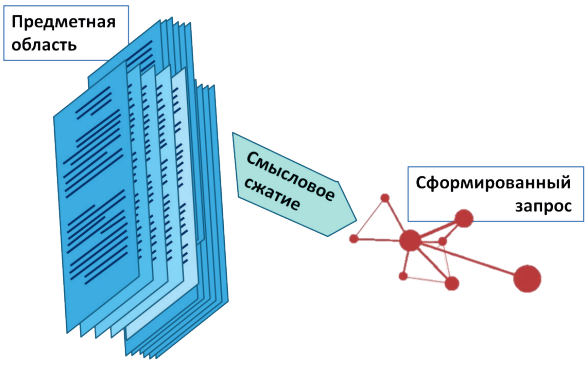

Рис. 3.

Именно поиск в Интернете сегодня является одной из наиболее часто решаемых задач, что редко имеют решения, которые бы нас полностью удовлетворяли. И прежде всего потому, что формирование запроса, равно и как сама система индексирования страниц не учитывает наших интересов. К тому же, алгоритмы обработки интернет запросов ориентированы на обеспечение релевантности, а не соответствие результатов нашим ожиданиям. Это связано не только с работой самих алгоритмов поиска и выдачи, но и формированием пользователем своих запросов, часто весьма далеких от точности (хотя понятие «точность» в данном контексте тоже весьма условно).

Использование семантических сетей, представляемых в виде слоеного текста, могут способствовать сокращению времени поиска, поскольку поиск осуществляется не по всей глубине имеющегося объема документов, а только по их значимой — главной — части (Рис. 4). Вполне возможно осуществлять предварительную обработку материалов, хранящихся на серверах для построения заранее семантических сетей, соответствующих хранимым данным.

Сравнение, прежде всего, частей структур, а только затем их содержания, может решительно изменить возможности поиска, как в Интернете, так и собственных документах, которые у пользователей хранятся на персональных компьютерах.

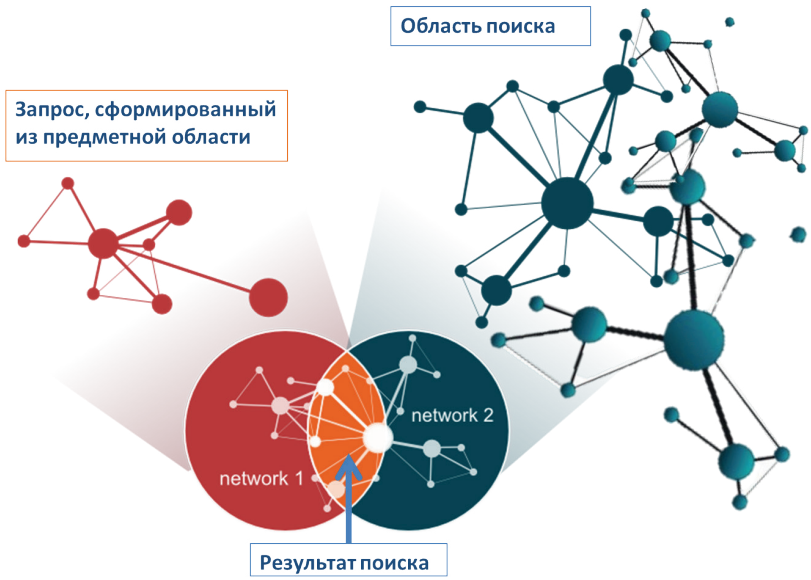

Рис. 4.

В заключение, хотелось бы отметить, что реферирование документов, традиционно осуществляемое людьми, нельзя отнести к семантическому сжатию, поскольку фактически это пересказ документов, что вносит субъективность в представление смысла, является трактовкой документа, в той или иной степени искажая его. Что недопустимо, если мы говорим о сжатии без потерь и искажений.

Я не нашел понятия, которое бы обозначало «смысловое сжатие без потери смысла». Потому что известные способы смыслового сжатия текста — исключение,

обобщение, упрощение — искажают исходный текст. К тому же, в нашем случае, речь идет скорее об изменении представления текста, а не уменьшении его объема. Поэтому я позволил себе предложить новое слово, отражающее суть данной технологии — семпрессия. Что значит — сем(антическая ком)прессия — смысловое сжатие, представление текста в сокращенном, сжатом виде, без искажения и потери какой бы-то ни было информации.