Посилання на статтю

Дымо А.Б. Программная реализация проекта системы автоматизированного реферирования / А.Б. Дымо, А.С. Морозова // Управління проектами та розвиток виробництва: Зб.наук.пр. – Луганськ: вид-во СНУ ім. В.Даля, 2011. – № 1(37). – С. 49-54. - Режим доступу: http://www.pmdp.org.ua/images/Journal/37/11dabsar.pdf

УДК 681.5

А.Б. Дымо, А.С. Морозова

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ ПРОЕКТА СИСТЕМЫ АВТОМАТИЗИРОВАННОГО РЕФЕРИРОВАНИЯ

Рассмотрены системы автоматизированного реферирования и решены проблемы увеличения их производительности в онлайновых поисковых системах с применением грамматик связей для синтаксического анализа и нейронных сетей для анализа структуры текста. Рис. 2, ист. 7.

Ключевые слова: системы автоматизированного реферирования, поисковые системы, нейронные сети.

О.Б. Димо, Г.С. Морозова

ПРОГРАМНА РЕАЛІЗАЦІЯ ПРОЕКТУ СИСТЕМИ АВТОМАТИЗОВАНОГО РЕФЕРУВАННЯ

Розглянуті системи автоматизованого реферування і вирішені проблеми збільшення їх продуктивності в онлайнових пошукових системах із застосуванням граматик зв'язків для синтаксичного аналізу і нейронних мереж для аналізу структури тексту. Рис. 2, дж. 7.

A.B. Dymo, A.S. Morozova

THE PROGRAM REALIZATION OF THE AUTOMATIC ABSTRACTING SYSTEM PROJECT

Systems of automatic abstracting are considered and problems of their productivity increasing in online search systems are solved using relationships grammatics for parsing and neuron networks for the text structure analysis.

Постановка проблемы. Реферирование текста лингвистическими методами в отличие от статистических на текущей архитектуре компьютеров зачастую является задачей, не решаемой за полиномиальное время [2]. Более того, многие задачи связанные с семантическим анализом текста при реферировании являются экспоненциальными, то есть невычислимыми за приемлемое время на текущей архитектуре.

Целью является поиск алгоритмов синтаксического и семантического анализа текста при реферировании с полиномиальной или лучшей производительностью.

Основной материал статьи. Несмотря на существенные проблемы реферирования с лингвистической точки зрения, проблемы программной реализации на первый взгляд носят характер акциденций, то есть сопутствующих трудностей. Такое рассмотрение аспектов программной реализации, несомненно, является верным, так как лингвистические задачи

являются решаемыми (вычислимыми). Однако, как будет показано далее, удачное технологическое обеспечение процесса реферирования может повлиять и на некоторые его существенные характеристики.

Для начала рассмотрим наиболее очевидные аспекты создаваемой системы. В целом, задача реферирования разбивается на подзадачи синтаксического анализа предложений, структурного анализа предложений и абзацев и компоновки.

Синтаксический анализ может быть выполнен с помощью контекстно-свободного анализатора Link [3], идея которого восходит к модели непосредственно составляющих Бархударова [2]. Независимость от контекста не представляется проблемой для задачи автоматизированного реферирования, так как, во-первых, грамматический центр предложений (тройка "субъект-предикат-объект") существует, так или иначе вне зависимости от контекста, а, во-вторых, поступающие на вход системы реферирования предложения могут считаться верными с точки зрения семантики, прагматики и связей реального мира. В то же время, нельзя не считаться со значительно большими показателями производительности контекстно-свободных грамматик по сравнению с контекстно-зависимыми. Так, согласно данным [2] производительность алгоритма Link – O(N3), где N – количество слов в предложении.

Структурный анализ включает в себя:

– сопоставление частей поступающих предложений шаблонам, соответствующим определенным структурам;

– сопоставление целых предложений и абзацев правилам согласованности.

Рассмотрение первой из задач структурного анализа приводит к выводу о том, что каждая группа шаблонов для некоторого элемента структуры Ei есть не что иное, как множество регулярных выражений [1], определяющих некий регулярный язык из предложений, соответствующих этому элементу Ei.

Программная реализация алгоритма сопоставления представляет собой недетерминированный конечный автомат (НКА) с ε – переходами с максимальным временем работы O(N2). Число состояний НКА M равно O(Rl), где Rl – суммарная длина регулярного выражения R, полученного дизъюнкцией выражений всех шаблонов всей структуры E. Ввиду того, что шаблоны могут быть модифицированы пользователем, регулярное выражение R должно преобразовываться в НКА после каждой модификации, можно утверждать, что время сопоставления в процессе структурного анализа не превысит O(N2*Rl).

Правила согласованности по структуре напоминают грамматику Link, что позволяет применить аналогичную технику анализа (исключая, этап лексического анализа), на этот раз на уровне предложений и целых абзацев. Очевидно, что анализ производится за время O(Sc3*Pc3), где Sc – количество предложений в тексте, а Pc – количество абзацев в тексте.

Компоновка предполагает сравнение грамматических центров предложений, соответствующих одному элементу структуры Ei. Само сравнение занимает линейное время. Также линейное время занимает нахождение деноминатора. Таким образом, общее время компоновки займет O(SPc), где SPc – количество пар предложений, подлежащих компоновке.

Как показано выше, проблема реферирования является для автоматизированной системы P – проблемой, т.е. разрешимой за полиномиальное время (в нашем случае за O(Pc*Sc2*N5*Rl*SPc)). Такая характеристика системы, безусловно, является важной. Однако большое количество определяющих параметров и немалые показатели степени снижают потенциальные области применения. Повышение же производительности

позволит применять предлагаемый метод реферирования в онлайновых поисковых системах, до сих пор, использующих менее адекватные, но более производительные статистические методы со временем работы O(Sc*N*SPc).

Представляется возможным подойти к проблеме увеличения производительности системы реферирования с трех сторон, соответствующих синтаксическому анализу, структурному анализу и компоновке. Однако наиболее очевидный способ лежит в сведении времени сопоставления предложения элементу структуры от O(N2*Rl) к O(N). Характеристика O(N) в данном случае является целью, так как она является скоростью работы человека на той - же задаче. Если, задаться целью, определить алгоритм сопоставления, в какой - то мере соответствующий процессу, происходящему в человеческом мозге, то можно увидеть путь, ведущий к такой цели. Наиболее удачными средствами моделирования работы мозга есть нейронные сети, описанные, например, в [4].

Нейронная сеть представлена весьма удачной аналогией человеческому мозгу применительно к задаче сопоставления предложений элементам структуры по двум причинам. Первая причина – это то, что, как и мозг, сеть в состоянии выдать состояния активации (сигналы) в выходном слое нейронов за время O(N) в случае последовательной подачи слов из предложений во входной слой и даже за время O(1) при подаче всего предложения целиком. Второй причиной применимости аналогии есть свойство сетей обучаться и аппроксимировать. Простая сеть, с прямой подачей (без реккурентности и самообучения), будучи обучена с помощью разработанной системы автоматизированного реферирования на некотором обучающем множестве предложений L, может затем выдавать заключения о предложениях Si не из множества L основываясь на “похожести” Si предложениям из множества L, то есть выполнять ту же эвристическую процедуру, что и человек в процессе реферирования.

Рассматривая современные топологии нейронных сетей, можно выделить два их вида, соответствующие двум возможным решениям задачи сопоставления.

Первое решение, оно же и наиболее очевидное, состоит в том, что для каждого элемента структуры Ei создается и обучается сеть Хопфильда [5], представленная на рис. 1.

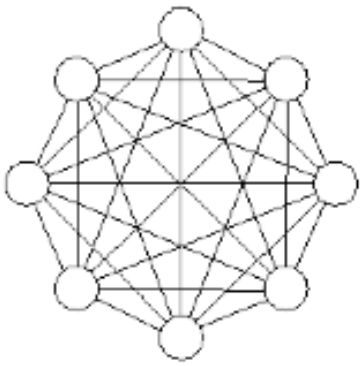

Рис. 1. Нейронная сеть Хопфильда

Так как в сети каждый нейрон является одновременно входным и выходным элементом, то слова предложения подаются на вход всех нейронов одновременно, а после обучения каждый нейрон в сети должен выдавать сигнал “1” для всех предложений обучающего множества L. В процессе работы такая сеть будет выдавать на всех нейронах сигнал либо “1” либо “0” в зависимости от

степени соответствия подаваемых предложений тому элементу структуры, для которого эта сеть была обучена. Очевидно, что потребуется столько сетей Хопфильда, сколько определено элементов структуры, что несколько снизит скорость сопоставления (до O(N*Ne), где Ne – количество элементов структуры).

Вторым решением будет применение самообучающейся сети Кохонена [6,7], представленной на рис. 2.

а б в г

а – необученная; б – после 200 - й итерации обучения;
в – после 600 - й итерации обучения; г – после 1900 - й итерации обучения

Рис. 2. Нейронная сеть Кохонена

В начале своего существования топология сети будет представляться множеством нейронов, соединенных между собой случайным образом и случайным – же образом активирующихся при подаче предложений на вход. После начального обучения сети нейроны будут организованы таким образом, что при подаче предложений, принадлежащих одной структуре, будут активироваться (выдавать сигнал “1”) только нейроны из некоторой, вполне определенной пространственной области. Таким образом, по принадлежности активированных нейронов областям будет определяться принадлежность предложения элементу структуры. Побочным, но, несомненно, полезным, эффектом будет самообучаемость сети, которая модифицируется при поступлении похожих предложений, не входящих в обучающее множество.

Оба обозначенных решения представляют несомненный интерес, однако оставляют обширное поле для дальнейших исследований. Так, нерешенным остался вопрос цифрового представления предложений, подаваемых на вход сети. Одним из возможных путей оцифровки предложения представляется группа методов, используемых конвертерами текст – голос, где текст аппроксимируется периодической функцией. Еще одним важным вопросом является изучение адекватности той “эвристики”, которая будет присуща обученной сети при сопоставлении предложений, не входящих в обучающее множество. Однако, такой анализ, представляет возможным в данное время, только экспериментальным.

Выводы. В работе доказано, что применение грамматик связей для реферирования текста решает проблему вычислимости. Этим способом автоматизированное реферирование будет решаться за время O(Pc*Sc2*N5*Rl*SPc), где N – количество слов в предложении, Sc – количество предложений в тексте, Pc – количество абзацев в тексте, Rl – суммарная длина регулярного выражения R, а SPc – количество пар предложений.

Также в работе показано, что применение нейронных сетей для семантического и структурного анализа текста позволит снизить время выполнения этих этапов реферирования вплоть до O(N*Ne), где Ne – количество элементов структуры.

ЛИТЕРАТУРА

1. Хопкрофт Дж. Введение в теорию автоматов, языков и вичислений / Дж. Хопкрофт, Р. Мотвани, Д. Ульман. – М.: Издательский дом “Вильямс”, 2002. – 528 с.
2. Бархударов Л.С. Структура простого предложения современного английского язика / Л.С. Бархударов. – М.: Издательство “Высшая школа”, 1966. – 200 с.
3. Sleator D.D. Parsing English with a Link Grammar / D.D. Sleator, D. Temperley. – Pittsburg: School of Computer Science, Carnegie Melon University, 1993. – 14 p.
4. Krose B. An introduction to Neural Networks / B. Krose, P. van der Smagt. – Amsterdam: The University of Amsterdam, 1996. – 135p.
5. Hopfield J.J. Neural networks and physical systems with emergent collective computational abilities / J.J. Hopfield //Proceedings of the National Academy of Sciences, №79, 1982. – pp. 2554-2558.
6. Kohonen T. Self-organized formation of topologically correct feature maps / T. Kohonen // Biological Cybernetics, №43. – 1982. – pp.59-69.
7. Kohonen T. Self-Organization and Associative Memory / T. Kohonen. – Berlin: Springer-Verlag. – 2001.

Рецензент: Кошкін К.В., професор, д.т.н.

Стаття надійшла до редакції 18.12.2010 р.