Система автоматического реферирования методом симметричного реферирования

Ступин Вячеслав Сергеевич

Вступление

Программа автоматического реферирования разработана с применением системы управления базами данных Visual FoxPro 6.0 (VFP). Даже такие современные языки программирования, как Visual C++ 6.0 или C#, имеют ограничения по объему памяти для текстовых переменных (например, строковая – 255 байт). VFP позволяет использовать базы данных (в частности несвязанные между собой таблицы) для обработки текстов больших объемов (например, размер таблицы до 2 ГБ [1]). Рационально будет использовать таблицу с текстом, разбитым по предложениям для последующей работы с ними (сам текст нужен как архив (копия) документа).
Программа реферирования реализована в автономном режиме (версия для работы с пользователем, Рис. 1), но с возможностью использования в качестве модуля АИПС (Рис. 2). При работе с пользователем программа работает с исходным текстом, месторасположение которого указывает пользователь, и тематическими словарями, находящимися локально и используемыми по указанию пользователем темы (категории). На выходе пользователю предоставляется готовый текст реферата с количеством предложений в соответствии с заданным пользователем значением, влияющем на количество предложений в реферате.
Отличие модульной версии заключается в получении программой от АИПС исходного текста программа вместе с категоризированным словарем и отправкой в АИПС текста реферата с заданным количеством предложений, а также разными вариантами количества предложений текста реферата.

Описание существующих аналогов

Разработан ряд автоматических информационно-поисковых системах (АИПС), систем аннотирования и реферирования: это такие инструменты, как функция AutoSummarize в Microsoft Office 97, системы IBM Intelligent Text Miner, Oracle Context и Inxight Summarizer (компонент АИПС AltaVista), безусловно, полезны, как и многие другие, но их возможности ограничены выделением и выбором оригинальных фрагментов из исходного документа и соединением их в короткий текст. Подготовка же краткого изложения предполагает передачу основной мысли текста, и не обязательно теми же словами. Текст, полученный путем соединения отрывочных фрагментов, лишен гладкости, его трудно читать. Кроме того, источники информации вовсе не всегда являются текстами, ведь необходимо подготавливать аннотации и на видеозаписи, к примеру, спортивных соревнований, или формировать сводные данные по биржевым таблицам. Перечисленные

инструменты реферирования рассчитаны на обработку только текстовой информации и не могут работать сразу с несколькими источниками.

Smart Search System (3S)

Проект 3S начат компанией InnerSpace [2] в конце 1999 года. К настоящему времени удалось создать систему, которая объединяет в себе высокую релевантность результатов (что свойственно системам, использующим глубокий семантический разбор текстов). В отличие от простого поиска по наличию в тексте тех или иных слов, алгоритм семантического поиска использует ассоциации между терминами, выявленные при структуризации коллекции документов. Возможности поиска расширяются за счет автоматического добавления в запрос связанных с этим запросом терминов, которые уже содержатся в семантической сети данной коллекции документов.

Inxight Summarizer

На рынке присутствует очень небольшое количество традиционных программ реферирования, то есть таких, которые выделяют наиболее весомые предложения из текста используя статистические, алгоритмы, либо слова-подсказки. Inxight Summarizer [3] — одна из наиболее известных коммерчески распространяемых систем реферирования. Inxight Summarizer был создан в Исследовательском центре Ксерокса в Пало Альто.

Extractor

Программа создана в Институте Информационных Технологий Национального исследовательского Совета Канады [3]. Он представляет собой модуль, выделяющий из представленного ему на вход текста наиболее информативные именные группы. По умолчанию количество таких групп — 7 вне зависимости от длины текста. Extractor используется в программных продуктах фирм ThinkTank Technologies и Tetranet, а также в поисковой системе Журнала Исследований в Области Искусственного Интеллекта.

TextAnalyst

Программа создана в Московском Научно-производственном Инновационном Центре «МикроСистемы» [3]. TextAnalyst работает только с русским языком, выделяя именные группы и строя на их основе семантическую сеть — структуру взаимозависимостей между именными группами.

Золотой ключик

Это программная библиотека, работающая по принципу фильтрации на базе тезауруса [4]. Как входные данные программе подается произвольный текст на русском языке, на стандартном выходе программа формирует аннотацию данного текста и список рубрик, к которым относится данный текст. В качестве аннотации используются предложения из входного текста, наиболее полно отражающие тематику текста. При рубрикации текста используется фиксированный список заранее определенных рубрик.

МЛ Аннотатор

Программа составляет связный реферат документа. Относительный размер реферата («коэффициент сжатия») задаётся пользователем. Программа имеет два режима работы: собственно реферирование и выделение ключевых слов. В режиме реферирования из текста отбираются предложения, в наибольшей степени характеризующие его содержание [5]. В режиме выделения ключевых слов производится выборка из текста наиболее информативных слов. Программа выделяет в тексте значимые и шумовые слова,

самостоятельные и зависимые предложения, определяет семантический вес предложений и удаляет незначащие фрагменты. Отобранные предложения при необходимости слегка перефразируются [6]. Используются специальные вероятностные модели, машинная морфология русского языка и другие интеллектуальные алгоритмы.
Существующие системы автореферирования являются дискретными, что дает стимул для создания непрерывной (многодокументных) системы автореферирования, что необходимо для обработки набора Internet-документов в базе данных поисковой системы. В Хакасском государственном университете им. Н.Ф. Катанова разработан метод симметричного реферирования [7] (основан на тематических словарях[8]), который позволяет применять систему автоматического непрерывного реферирования в совокупности с автоматическими информационно-поисковыми системами для обеспечения достаточной точности и релевантности поиска.

Описание разработанной программы

В программе используется словарь из ключевых слов книги [9] по теме Windows NT, что определяет направленность работы программы с текстами, ориентированными на тему Windows NT. В дальнейшем будет добавлена возможность работы со словарями других тематик. Тематический словарь (рис. 3) имеет структуру, состоящую из пронумерованного списка тематических слов, имеющих признаки наличия флективных форм и гипонимов (термин является гиперонимом), номер гиперонима (термин является гипонимом).
Архитектура программы представлена на рис. 4.
Тематический словарь (таб. 1) представляет собой нумерованный список ключевых слов по категории Windows NT. Нумерация не связано с алфавитным порядком, так как при дополнении новых слов могут нарушиться связи с гипонимами. Предполагается, что длины поля для ключевых слов размером в 50 символов будет достаточно, а количество слов может достигать 99999. Связь гипонимов с гиперонимами осуществляется по номеру гиперонима, хранимого в записи гипонима.
Таблица флективных форм отсутствует – для уменьшения объема словаря возможность проверки флективных форм на наличие в исходном тексте реализована в виде функции генерирования флективных форм (множественное число, отглагольное существительное) ключевых слов на основе признака флективности (наличия флективных форм) термина. Встроена эта функция в процесс проверки текста на наличие словарных терминов в предложениях.
Исходный текст, текстовый файл в формате HTML, загружается в таблицу с Memo-полем для его хранения там и дальнейшем обработки. В модульном режиме АИПС поставляет для реферирования образ исходного текста. Изначально программа определяет в загруженном тексте (со структурой заголовков) местонахождения тэгов заголовков (h1, /h1, h2 /h2 и другие) и сохраняет их в таблице (список заголовков).
Этот список предполагается использовать для создания тематического словаря ключевых слов для программы автоматического реферирования. Заголовки выбираются по позициям тэгов заголовков во всем тексте, затем заголовки выбираются в таблицу на сохранение. В модульном режиме возможна категоризация текста и определение тематических словарей со стороны АИПС.
Необходимо перед дальнейшей обработкой текста очистить его от дублирующих символов 0Dh и 0Ah (перевод строки и возврат каретки), так как при создании текста HTML-генераторами или написанием кодами программистами эти символы изобилуют в тексте.

Обработка оставшейся части текста начинается с создания списка знаков препинания, ограничивающих предложения (‘.’, ’!’, ’?’). Все позиции знаков препинания запоминаются в таблице, что позволяет по этим записям выбирать предложения текста в таблицу предложений с их автоматической нумерацией. Также сохраняется порядок предложений в выходном тексте (реферате).

Обработка предложений ранее была организована по следующей схеме: а) поиск позиций знаков препинания и пробелов, б) выборка всех слов из предложений, в) сравнение их со списком ключевых слов. Последняя реализации программы сократила этот процесс вместе с затрачиваемым на него машинным временем до 30% от общего процесса реферирования. Это было реализовано за счет замены всех трех пунктов одним: сравнение символов предложений с символами ключевых слов. Например, следующая SQL-выборка (рис. 3) выбирает из таблицы предложений (temp_snt) все предложения, где есть в наличие комбинация символов (слово) термина ‘windows’.

Поиск ключевых слов в предложениях происходит с автоматической проверкой в предложении флективных форм ключевого слова, то есть проверяется наличие либо ключевого слова, либо его флективной формы. При наличии искомых последовательностей символов (ключевых слов) в предложениях эти предложения выбираются в таблицу найденных слов (таб. 2): номер предложения, номер слова.

Процесс генерации выходного текста происходит на основе полученного списка наличия ключевых слов в предложениях [7]. Подсчитывается количество вхождений слов в других предложениях как справа от текущего предложения, так и слева (метод симметричного реферирования). Также происходит подсчет весов предложений, содержащих флективные формы терминов, кореферентные термины и гипонимы терминов.

Для вывода реферата необходимо определить функциональную значимость предложений: K = sum_rel / sum_sent (коэффициент значимости = количество связей в предложениях / общее количество предложений). Отбираются в реферат все предложения с количеством связей не менее K.

Кореферентные слова отбираются при условии, что предложения в исходном тексте имеют коэффициент полезности текста (необходимого для реферирования кореферентных терминов) отличный от нуля: R = (sum_relsent * 100%) / sum_sent (коэффициент полезности текста = (количество предложений со связями * 100%) / общее количество предложений). Обработка предложений с кореферентными словами происходит при условии наличия ключевых слов и количеством связей предложения не менее R (relates => R), где предложения являются «полезнымы» для кореферирования.

Для подходящих под условие «полезности» гиперонимов выбираются гипонимы в таблицу гипонимов, схожей по структуре с тематическим словарем.

Процесс определения связей в предложениях схож с работой с ключевыми словами, только без флективных форм, так как нет признака наличия флективных форм у найденных гипонимов. Количества найденных левых и правых связей суммируются с найденными ранее связями.

Выходной текст генерируется (нахождение K и вывод реферата, состоящего из предложений с количеством записей не менее K) на экран либо в результирующий файл с сохранением порядка следований предложений в исходном тексте. В модульном режиме результирующая таблица, содержащая Memo-поле с рефератом, отправляется в АИПС с сохранением порядка следований предложений в исходном тексте. Дополнительные варианты, которые состоят из разного количества предложений реферата (3, 5, 10), также отправляются в АИПС.

Описание метода симметричного реферирования

Методика составления тематического словаря может включать следующие этапы и процедуры:

1. отбор лексики из заглавий статей (создается список всех использованных слов в заглавиях);
2. элиминация: а) служебных слов (артиклей и предлогов), б) прилагательных, используемых в качестве определений, с) повторяющихся слов;
3. выделение тематической лексики и нетематической лексики.

Приведем пример составления тематического словаря, на примере научного издания по теме «Windows NT Unleashed» [10], который основывается на списке ключевых слов оглавления статьи (П.1). После обработки словаря тема «Windows NT 4 Server Unleashed » включает следующие слова и термины (таб. N1)

Таблица N1. Заполненный тематический словарь.

Id_word	Words	Flexy	Hyper	Hypo
	Nos	False		
	Os	False		
	Dos	False		
	Windows	False		True
	Nt	False	4	
	16-bit	False		
	32-bit	False		True
	Operating	False	4	
	System	True		True
	Server	True	4	
	Client/server	False		
	Flat	True		
	Memory	True		
	Model	True		
	Protected	False		
	Preemptive	True		
	Multitasking	False		
	Portability	True		
	Scalability	True		
	Personality	True		
	Compatibility	True		
	Localization	True		
	Security	True		
	Fault-tolerance	False		
	Network	True	4	
	Net	True		
	Networking	False		
	Microsoft	False	4	

	200	False	4	
	File	True	4	
	Installing	False	4	
	Wins	False	4	
	Dhcp	False	4	
	Internet	False	4	
	Kernel	True	9	
	64-bit	False	7	

Согласно методике симметрического экстрагирования флективные формы этих слов тоже учитываются. Эти формы включают существительные во множественном числе (например, systems) и производные (например, protector), в таблице они находятся без признака флективности, как аббревиатуры и другие многословообразованные термины и сокращения. При реферировании больших по объёму текстов можно не учитывать атрибутивные существительные.

В тематический словарь также включены кореферентные слова (*networking = net = networks*) и гипонимы наиболее часто повторяющихся слов (*windows – microsoft*, *2000*, *file*, *installing*, *wins*, *dhcp*, *internet; system – kernel; 32-bit – 64-bit*).

Исходный текст разбирается по предложениям, которые содержат термины из тематического словаря (таб. N2).

Таблица N2. Пример таблицы предложений с терминами.

1.	If you ask about the differences between a 16-bit and a 32-bit operating system, you'll get all kinds of responses.
3.	The essential difference between 16-bit and 32-bit operating systems is the way they handle internal structures.
5.	Recently Microsoft announced they would work on a pseudo-64-bit version of NT.
6.	Although the new system will not be fully 64-bit, it will permit 64-bit data structures and a 64-bit flat memory space.
8.	Using 64-bit addressing, you can support extremely large databases.
10.	In fact, when we're talking about Microsoft Windows operating systems, there is a big difference between 16-bit and 32-bit versions.
11.	For example, one of the features of the 32-bit Microsoft operating systems, is support for a 32-bit protected, flat memory model, which provides cleaner memory management than 16-bit Windows, and allows programs to create and address very large data structures.
12.	The base of 32-bit Windows operating systems is a complete 32-bit kernel.
13.	The kernel does things such as system scheduling and memory management.

Симметричное реферирование основано на установлении весов (количества связей между предложениями). При первом проходе устанавливаются связи с правой стороны, при втором – с левой. Можно принимать во внимание только левосторонние связи, то есть связи с предшествующим текстом, не учитывая правосторонние связи, но в этом случае нарушается принцип симметричности, и реферирование становится асимметричным. За исходный текст будут взяты блоки *16-bit and 32-bit Operating Systems* и *Network Operating Systems* статьи *Introducing Windows NT* [10]. В исходном тексте находим веса предложений (таб. 4).

Итого, 556 связей между предложениями.

При использовании гипонимов словарных терминов *windows* (*microsoft*), *system (kernel)* и *32-bit (64-bit)* получается увеличение количества связей у предложений, что приводит к увеличению функциональных весов предложений, так как увеличивается вероятность нахождения кореферентных терминов и гипонимов в тексте. При использовании дополнительных гипонимов словарных терминов *nt, operating, network, server* (*windows*) увеличивается количество связей у предложений, что приводит также к увеличению функциональных весов предложений (на основе используемой структуры тематического словаря гипероним может являться гипонимом). Для реферата (выходного текста) необходимо указать его размер, который вычисляется из отношения количества связей к количеству предложений. Следовательно, это приводит к возрастанию информативности реферата:
$K = 556 / 25 = 22.24 \approx 22$.
Получается реферат следующего содержания (П. 2):
1, 3, 10, 11, 12, 15, 18, 23, 24, 25
С каждым увеличением числа гипонимов, найденных в тексте, происходит увеличение количества предложений, которые будут выведены в реферате исходного текста. В реферате окажутся следующие самые информативные предложения (таб. 5, 6):
- 10, 11, 12 (для варианта с 3 предложениями в реферате);
- 1, 10, 11, 12, 15 (для варианта с 5 предложениями в реферате);
- 1, 3, 10, 11, 12, 15, 18, 23, 24, 25 (для варианта с 10 предложениями в реферате).
В итоге, реферат имеет семантическую структуру, отличную от исходного текста, но с сохранением порядка предложений. Разработка метода оценки эффективности семантической структуры реферата будет проведена на следующих этапах научной работы.

Таблицы быстродействия программы

Для определения быстродействия программы были использованы файлы [10, 11] с размерами 23601 и 64142 байт соответственно, что позволило объективно оценить зависимость скорости процесса реферирования от процессора и объема оперативной памяти (ОЗУ). Результаты тестирования приведены в таблицах 7 и 8. Из протестированных конфигураций наиболее быстродействующей является под номером 1. Но на сегодняшний день самым оптимальным вариантом (по быстродействию) является следующая конфигурация – процессор Pentium III 650 Mhz и 128 MB ОЗУ.

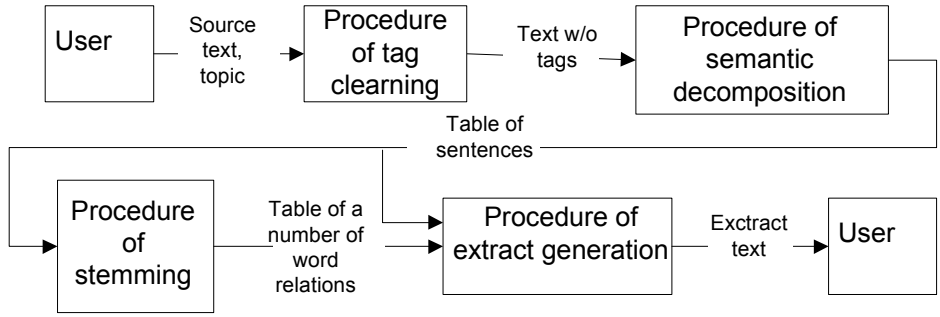
Таблицы сравнения программ автоместического реферирования

Для сравнительного анализа оппонирующей программой был выбран модуль «Автореферирование» из текстового процессора Word пакета MS Office XP. В качестве тестовых наборов выступали два текста [10, 11]. В задачу тестирования входило получение реферата из 10-ти предложений. Подученные результаты можно наблюдать в таблицах 11, 12.

Выводы

Разработанный В.А. Яцко метод симметричного реферирования позволяет создавать рефераты, информативные по содержанию. Программа, созданная на основе этого метода генерирует рефераты по своей структуре соответствующие структуре исходного текста. Процесс реферирования происходит медленнее, чем в Word (там исходный текст уже

разбит по предложениям), но не происходит разделения текста на заголовки и остальной текст. Программа FASS на основе разбора текста на заголовки и остальной текст создает более информативный реферат, чем Word.



Рисунки и таблицы

Рис. 1. Схема автономной работы программы реферирования.

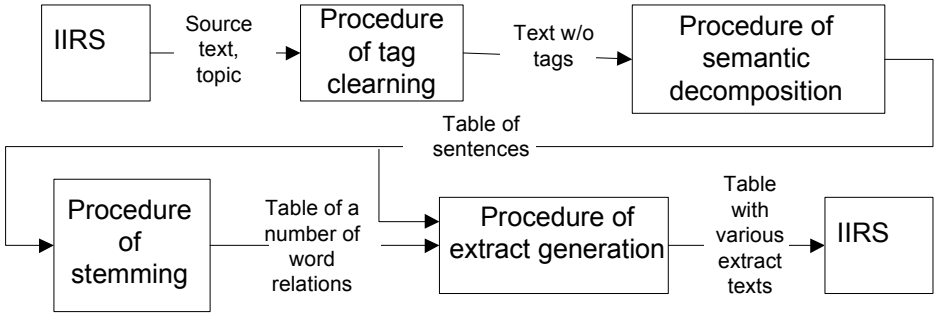


Рис. 2. Схема работы программы реферирования в качестве модуля АИПС.

```
word_1 = 'windows'
SELECT idsent, sent;
FROM temp_snt;
     WHERE (AT(word_1, sent)<>0
```

Рис. 3. SQL-выборка предложений с наличием словарного термина.

Таблица 1. Схема тематического словаря.

Таблица слов	**Тип**	**Длина**	**Комментарии**
Id_word	Число	4	Номер слова
Words	Строка	50	Слово
Flexy	Булево	1	Признак наличия флективных форм
Hyper	Число	4	= id_word (номер гиперонима)
Hypo	Булево	1	Признак наличия гипонимов

Таблица 2. Структура таблицы найденных слов.

Таблица найденных слов	**Тип**	**Длина**	**Комментарии**
Idsent	Число	4	= idsent (номер предложения из таблиц предложений)
Idword	Число	4	= idword (найденное слово из тематического словаря)

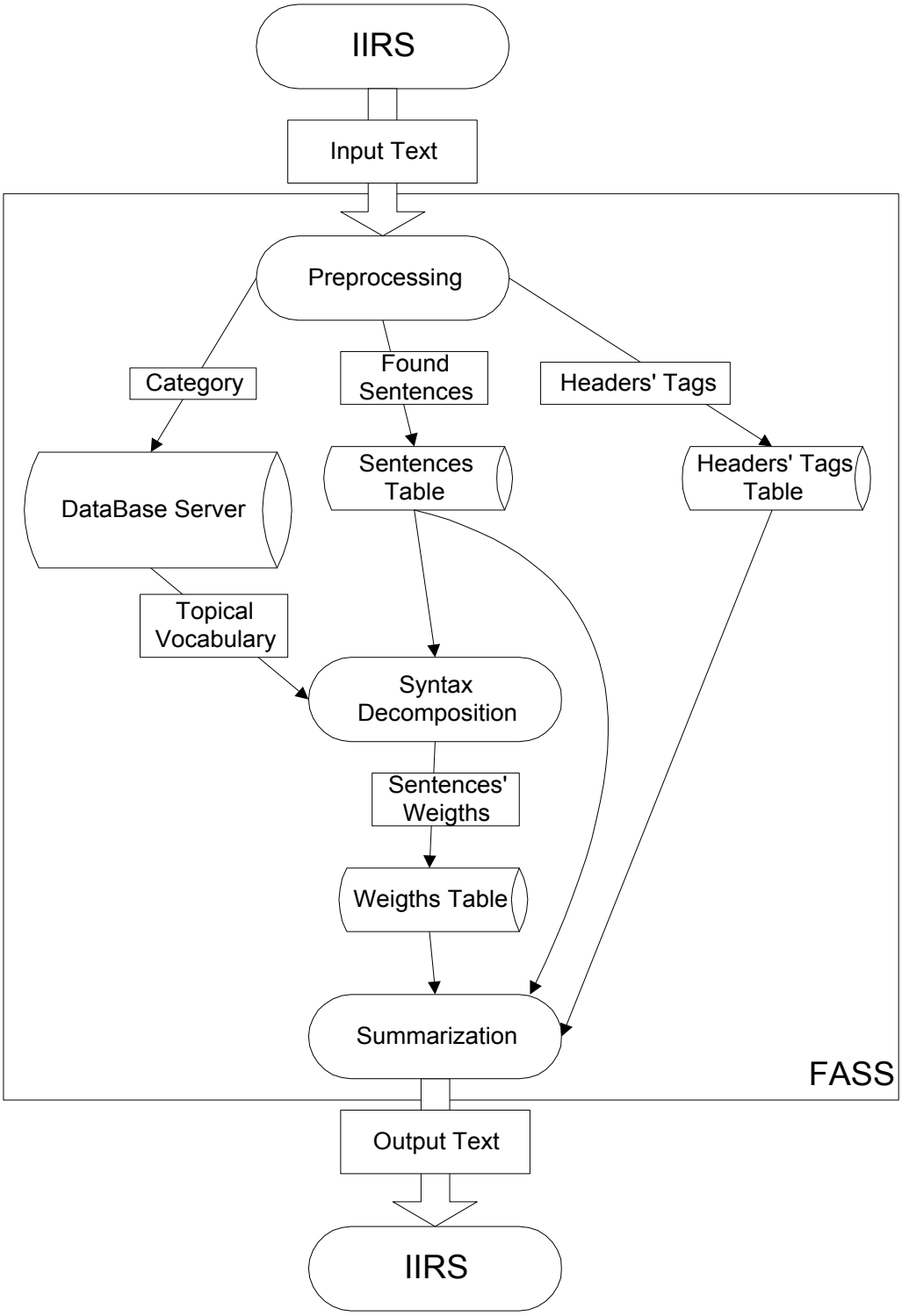


Рис. 4. Архитектура программы.

Таблица 4. Право- и левосторонние связи между предложениями.

№ пред.	Словарный термин	№ правого предложения, имеющего связь	Кол-во правых связей	№ левого предложения, имеющего связь	Кол-во левых связей	Все-го
1	16-bit	3, 10, 11	3			31
	32-bit	3, 10, 11, 12, 14, 15, 16, 17	8			
	Operating	3, 10, 11, 12, 18, 23, 25	7			

	System	3, 6, 10, 11, 12, 13, 14, 15, 17, 18, 20, 23, 25	13			
	Operating (windows)	3, 5, 10, 11, 12, 14, 15, 16, 18, 21, 23, 24, 25	13			13
3	16-bit	10, 11	2	1	1	
	32-bit	10, 11, 12, 14, 15, 16, 17	7	1	1	
	Operating	10, 11, 12, 18, 23, 25	6	1	1	31
	System	6, 10, 11, 12, 13, 14, 15, 17, 18, 20, 23, 25	12	1	1	
	Operating (windows)	5, 10, 11, 12, 14, 15, 16, 18, 21, 23, 24, 25	12	1	1	13
5	NT	16, 24	2			2
	NT(windows)	10, 11, 12, 14, 15, 16, 18, 21, 23, 24, 25	11	1, 3	2	13
6	System	10, 11, 12, 13, 14, 15, 17, 18, 20, 23, 25	11	1, 3	2	13
	64-bit (32-bit)	8	1			1
8	64-bit (32-bit)			6	1	1
10	Windows	11, 12, 15, 16, 24	5			
	Operating	11, 12, 18, 23, 25	5	1, 3	2	
	System	11, 12, 13, 14, 15, 17, 18, 20, 23, 25	10	1, 3	2	35
	16-bit	11	1	1, 3	2	
	32-bit	11, 12, 14, 15, 16, 17	6	1, 3	2	
	Microsoft (windows)	23	1	5	1	2
	Operating (windows)	11, 12, 14, 15, 16, 18, 21, 23, 24, 25	10	1, 3, 5	3	13
11	32-bit	12, 14, 15, 16, 17	5	1, 3, 10	3	
	Operating	12, 18, 23, 25	4	1, 3, 10	3	
	System	12, 13, 14, 15, 17, 18, 20, 23, 25	9	1, 3, 6, 10	4	34
	Memory	13	1			
	Windows	12, 15, 16, 24	4	10	1	
	Memory	13	1			
	Operating (windows)	12, 14, 15, 16, 18, 21, 23, 24, 25	9	1, 3, 5, 10	4	13
12	32-bit	14, 15, 16, 17	4	1, 3, 10, 11	4	
	Windows	15, 16, 24	3	11, 10	2	33
	Operating	18, 23, 25	3	1, 3, 10, 11	4	
	System	13, 14, 15, 17, 18, 20, 23, 25	8	1, 3, 6, 10, 11	5	
	Kernel (system)	13, 16	2			2
	Operating (windows)	14, 15, 16, 18, 21, 23, 24, 25	8	1, 3, 5, 10, 11	5	13
13	System	14, 15, 17, 18, 20, 23, 25	7	1, 3, 6, 10, 11, 12	6	14
	Memory			11	1	
	Kernel (system)	16	1	12	1	2
14	32-bit	15, 16, 17	3	1, 3, 10, 11, 12	5	8
	Operating (windows)	15, 16, 18, 21, 23, 24, 25	7	1, 3, 5, 10, 11, 12	6	13
15	32-bit	16, 17	2	1, 3, 10, 11, 12, 14	6	
	Windows	16, 24	2	10, 11, 12	3	33
	Operating	18, 23, 25	3	1, 3, 10, 11, 12	5	
	System	17, 18, 20, 23, 25	5	1, 3, 6, 10, 11, 12, 13	7	
	Operating (windows)	16, 18, 21, 23, 24, 25	6	1, 3, 5, 10, 11, 12, 14	7	13
16	32-bit	17	1	1, 3, 10, 11, 12, 14, 15	7	
	Windows	24	1	10, 11, 12, 15	4	15
	NT	24	1	5	1	
	Kernel (system)			12, 13	2	2
	NT(windows)	18, 21, 23, 24, 25	5	1, 3, 5, 10, 11, 12, 14, 15	8	13
17	Systems	18, 20, 23, 25	4	1, 3, 6, 10, 11, 12, 13, 15	8	12
18	Network	23, 24, 25	3			24

	Operating	23, 25	2	1, 3, 10, 11, 12, 15	6	
	System	20, 23, 25	3	1, 3, 6, 10, 11, 12, 13, 15, 17	9	
	Server	21	1			
	Operating, network (windows)	21, 23, 24, 25	4	1, 3, 5, 10, 11, 12, 14, 15, 16	9	13
20	Systems	23, 25	2	1, 3, 6, 10, 11, 12, 13, 15, 17, 18	10	12
21	Server (windows)	21, 23, 24, 25	3	1, 3, 5, 10, 11, 12, 14, 15, 16, 18	10	13
23	Operating	25	1	1, 3, 10, 11, 12, 15, 18	7	
	System	23, 25	2	1, 3, 6, 10, 11, 12, 13, 15, 17, 18, 20	11	24
	Network	24, 25	2	18	1	
	Microsoft (windows)			5, 10	2	2
	Operating, network (windows)	24, 25	2	1, 3, 5, 10, 11, 12, 14, 15	8	13
24	Windows			10, 11, 12, 15	4	6
	NT			5, 16	2	
	NT(windows)	25	1	1, 3, 5, 10, 11, 12, 14, 15, 16, 18, 21, 23	12	13
	NT(windows)	25	1	1, 3, 5, 10, 11, 12, 14, 15, 16, 18, 21, 23	12	13
25	Network			18, 23	2	
	Operating			1, 3, 10, 11, 12, 15, 18, 23	8	22
	System			1, 3, 6, 10, 11, 12, 13, 15, 17, 18, 20, 23	12	
	Operating, network (windows)			1, 3, 5, 10, 11, 12, 14, 15, 16, 18, 21, 23, 24	13	13

Таблица 5.Функциональные веса предложений.

№ пред.	Итого
1	44
3	44
5	15
6	14
8	1
10	50
11	47
12	48
13	16
14	21
15	46
16	30
17	12
18	37
20	12
21	13
23	39
24	32
25	35

Таблица 6.Функциональные веса предложений в порядке уменьшения.

№ пред.	Итого
10	50
12	48
11	47
15	46
1	44
3	44
23	39
18	37
25	35
24	32
16	30
14	21
13	16
5	15
6	14
21	13
17	12
20	12
8	1

Таблица 7. Характеристики тестированных систем для первого тестового файла.

Место	Процессор	ОЗУ	Операционная система	Время
	Pentium III 650 Mhz	128 MB	Windows 2000 Pro SP 1	0:26
	Celeron 366 Mhz	32 MB	Windows 95 OSR 2	0:40
	Celeron 366 Mhz	32 MB	Windows 98	0:57
	AMD K6-266 Mhz	32 MB	Windows 98	1:36
	Pentium-MMX 166 Mhz	16 MB	Windows 95 OSR 2	1:37
	Pentium 100 Mhz	32 MB	Windows NT 4 WS SP 5	2:00
	Pentium 100 Mhz	32 MB	Windows 2000 Pro SP 1	2:12
	Pentium 100 Mhz	32 MB	Windows 95 OSR 2	2:39

Таблица 8. Характеристики тестированных систем для второго тестового файла.

Место	Процессор	ОЗУ	Операционная система	Время
	Pentium III 650 Mhz	128 MB	Windows 2000 Pro SP 1	1:16
	Celeron 366 Mhz	32 MB	Windows 95 OSR 2	3:04
	Celeron 366 Mhz	32 MB	Windows 98	4:01
	Pentium-MMX 166 Mhz	16 MB	Windows 95 OSR 2	5:27
	AMD K6-266 Mhz	32 MB	Windows 98	5:58
	Pentium 100 Mhz	32 MB	Windows NT 4 WS SP 5	6:55
	Pentium 100 Mhz	32 MB	Windows 2000 Pro SP 1	7:20
	Pentium 100 Mhz	32 MB	Windows 95 OSR 2	10:24

Таблица 9. Сравнение рефератов.

№ предложения	«Автореферирование» Word из пакета MS Office XP	Summarize 0.8
	Introducing Windows NT	Windows NT Server 4 Unleashed If you ask about the differences between a 16-bit and a 32-bit operating system, you'll get all kinds of response
	What is Windows NT	NOTE:Recently Microsoft announced they would work on a pseudo-64-bit version of NT.
	Windows NT contains no DOS code in the operating system.	In fact, when we're talking about Microsoft Windows operating systems, there is a big difference between 16-bit and 32-bit versions.
	Design Objectives of Windows NT Server	For example, one of the features of the 32-bit Microsoft operating systems, is support for a 32-bit protected, flat memory model, which provides cleaner memory management than 16-bit Windows, and allows programs to create and address very large data structures.
	Client/Server Operating System	The base of 32-bit Windows operating systems is a complete 32-bit kernel.
	DOS was designed as a 16-bit operating system. *All* the operating system code in Windows NT is re-entrant.	Additionally, the 32-bit OS enables us to use 32-bit device drivers, which, among other advantages, enable the operating system to communicate with devices faster.
	Windows 95 also supports Unicode.	Most of the other features that come from 32-bit Windows operating systems come from their support of the Win32 API.
	Network Operating Systems	This API set can only be fully implemented on a 32-bit kernel, such as Windows NT and Windows 95.
	Windows NT is both an operating system and a network operating system.	No More DOS Perhaps one of the greatest accomplishments for Windows NT was to get rid of DOS completely.
		In fact, when Microsoft first began work on the NT project, there were no firm plans to enable NT to run DOS or Windows applications.

Таблица 10. Сравнение рефератов.

№ предложения	«Автореферирование» Word из пакета MS Office XP	Summarize 0.8

	Features Common to both Windows NT Server and Windows NT Workstation	When Microsoft introduced Windows NT in 1993, they offered two products: Windows NT 3.1 and Windows NT Advanced Server 3.1.
	Additional Features in Windows NT Server	With the introduction of 3.5 in late 1994, Microsoft changed the product names, their feature sets, and gave a clear indication of what roles each product was designed for.
	Optimization of NT Server and NT Workstation	Windows NT became Windows NT Workstation, and Windows NT Advanced Server became Windows NT Server.
	Windows NT Workstation Versus Windows NT Server	Windows NT Workstation was designed as a robust, 32-bit multithreaded, multitasking operating system that was capable of running high-end engineering or mission-critical client/server applications.
	Windows NT became Windows NT Workstation, and Windows NT Advanced Server became Windows NT Server.	Windows NT Server became the cornerstone of Microsoft’s enterprise-class network operating system.
	Features Common to both Windows NT Server and Windows NT Workstation	Windows NT Server was designed to provide file, print, and application services to diverse clients.
	NT File System (NTFS)	Features Common to both Windows NT Server and Windows NT Workstation Windows NT Workstation and Windows NT Server are both built using the same core technologies, resulting in products with more similarities than differences.
	Additional Features in Windows NT Server	The Windows NT platform was designed to provide a powerful operating system platform capable of scaling from the simplest file and print services network, to the largest enterprise network providing file and print services to thousands of users, as well as advanced messaging and application services.
	Windows 95	To achieve this, Windows NT was designed with a microkernel capable of preemptively dispatching threads to up to 32 processors.
	Optimization of NT Server and NT Workstation	Windows NT is a true 32-bit operating system, with no internal 16-bit code, unlike Windows 95, which still has a considerable amount of 16-bit code under the hood for compatibility with older versions of Windows.

Литература

1. Microsoft Corp. Visual FoxPro Help (foxhelp.hlp)
2. InnerSpace Company, Smart Search System , 2001. http://www.innerspace.ru
3. А. Михаилян. Некоторые методы автоматического анализа естественного языка, используемые в промышленных продуктах, 2000. http://www.inteltec.ru/publish/articles/textan/natlang.shtml
4. Textar. Golden Key, 1998. http://www.textar.ru/gkey.html
5. ЗАО «Компания CPS». Либретто, 2001. http://www.cps.ru/vendors_ru/medialing/libretto.shtml
6. SoftLine. МЛ Аннотатор 1. 0 для Windows 95 и Windows NT, 2001. http://www.softline.ru/products/MediaLingua/MLAnnotator/MLAnnotator1Win_full.asp
7. Iatsko V. Linguistic Aspects of Summarization// Philologie in Netz – №18/2001. – s. 33-46. http://www.fu-berlin.de/phin/phin18/p18t3.htm
8. Лукашевич Н.В. Автоматизированное формирование информационно-поискового тезауруса по общественно-политической жизни России // НТИ. Сер.2. - 1995. - N 3. - С.21-24.
9. Ресурсы Microsoft Windows NT Workstation 4.0, под редакцией Екатерины Кондуковой: пер. с англ. – СПб.: BHV – Санкт-Петербург, 1998. – 800 с.: ил.
10. Microsoft Corp. Introducing Windows NT. http://www.mcp.com/846201600/0-672/0-672-30933-5/ch01.htm

11. Microsoft Corp. Introducing Windows NT. http://www.mcp.com/846201600/0-672/0-672-30933-5/ch03.htm

Приложение

П.1 Ключевые (заголовки) слова содержания статьи *Introducing Windows NT*
NOS, OS, But No DOS: An Introduction to Windows NT; What is Windows NT; 16-bit and 32-bit Operating Systems; No More DOS; Design Objectives of Windows NT Server; Client/Server Operating System; Flat, 32-bit Memory Model; Reliability Through Protected Memory Model; Preemptive Multitasking; Portability; Scalability; Personality/Compatibility; Localization; Security; Fault-Tolerance; Network Operating Systems; What is a Network Operating System?; Summary

П.2 Текст реферата из 10 предложений статьи *Introducing Windows NT*
1. If you ask about the differences between a 16-bit and a 32-bit operating system, you'll get all kinds of responses. 3. The essential difference between 16-bit and 32-bit operating systems is the way they handle internal structures. 10. In fact, when we're talking about Microsoft Windows operating systems, there is a big difference between 16-bit and 32-bit versions. 11. For example, one of the features of the 32-bit Microsoft operating systems, is support for a 32-bit protected, flat memory model, which provides cleaner memory management than 16-bit Windows, and allows programs to create and address very large data structures. 12. The base of 32-bit Windows operating systems is a complete 32-bit kernel. 15. Most of the other features that come from 32-bit Windows operating systems come from their support of the Win32 API.
18. A network operating system has traditionally been a method for describing the methods and protocols used by network clients when communicating with a network server. 23. Most of Microsoft's network-related products have obscured the line between operating system and network operating system. 24. Windows NT is definitely no exception to this. 25. As we move into a period in computing when the network becomes more and more an integral part of the entire environment, the concept of and the need for a network operating system will disappear or, if you prefer, simply swallow up the underlying standalone operating system.