ПРОБЛЕМЫ ОРГАНИЗАЦИИ ЭЛЕКТРОННОГО АРХИВА С СЕМАНТИЧЕСКИМ ИНДЕКСИРОВАНИЕМ ДОКУМЕНТОВ

И.С. Кононенко, Е.А. Сидорова, Ю.А. Загорулько, Ю.В. Костов
Российский НИИ Искусственного Интеллекта,
Институт систем информатики СО РАН
irina@mail.nsk.ru , lena@iis.nsk.su

Введение

Необходимой функцией корпоративной информационной системы является обеспечение управления потоком входящих документов. Система InDoc, разрабатываемая коллективом РосНИИ ИИ, решает две основные задачи: а) автоматическая классификация и оперативное распределение входящих документов среди сотрудников организации (см. [1]); б) передача документов в электронный архив и поиск в нем. В данном сообщении рассматриваются основные аспекты создания электронного архива в рамках системы InDoc: подход к индексированию и поиску в архиве, поддержка актуальности индекса и повышение эффективности работы с архивом.

В традиционных системах поиска учитываются поверхностные характеристики текстов, такие как ключевые слова и статистические оценки их распределения [2]. Как неоднократно отмечалось, применение методов полнотекстового поиска (простого или расширенного, с использованием морфологии) позволяет добиться высокой скорости обработки текста, но приводит к потере полноты и точности. В последние годы предлагается применение тезаурусов для автоматического концептуального индексирования [3], что дает возможность уточнения запросов и расширения поиска на основе тезаурусных связей. Высокие требования, предъявляемые к качеству поиска документов в архиве, предполагают привлечение более глубоких методов анализа содержания с выявлением семантических связей понятий в составе высказываний и распознаванием ситуаций в духе задачи извлечения информации (подобный подход обсуждается в [4]). При этом выбор лингвистических средств обработки должен быть достаточно гибким, учитывать специфику подъязыка документов и структуру предметной области. Наличие в базе знаний системы InDoc структурированной информации о предметной области позволяет отказаться от традиционных методов полнотекстового поиска и применить технологию семантического индексирования.

1. Общая схема функционирования системы InDoc

Работа с документами в системе включает три контура:

- ввод и первичная обработка документов,
- автоматическая обработка, индексирование, рассылка и архивирование документов,
- оперативный поиск и выдача документов.

Документ поступает на вход системы (Рис. 1.) через модуль ввода. С каждым документом в архиве ассоциируется ***электронный паспорт.*** Все атрибуты электронного паспорта разделены на служебный раздел и индекс; к индексу относятся атрибуты паспорта, по которым может осуществляться навигация и поиск документа в архиве. Набор атрибутов, связанных с содержанием документа, составляет ***семантический индекс***, который заполняется системой автоматически.

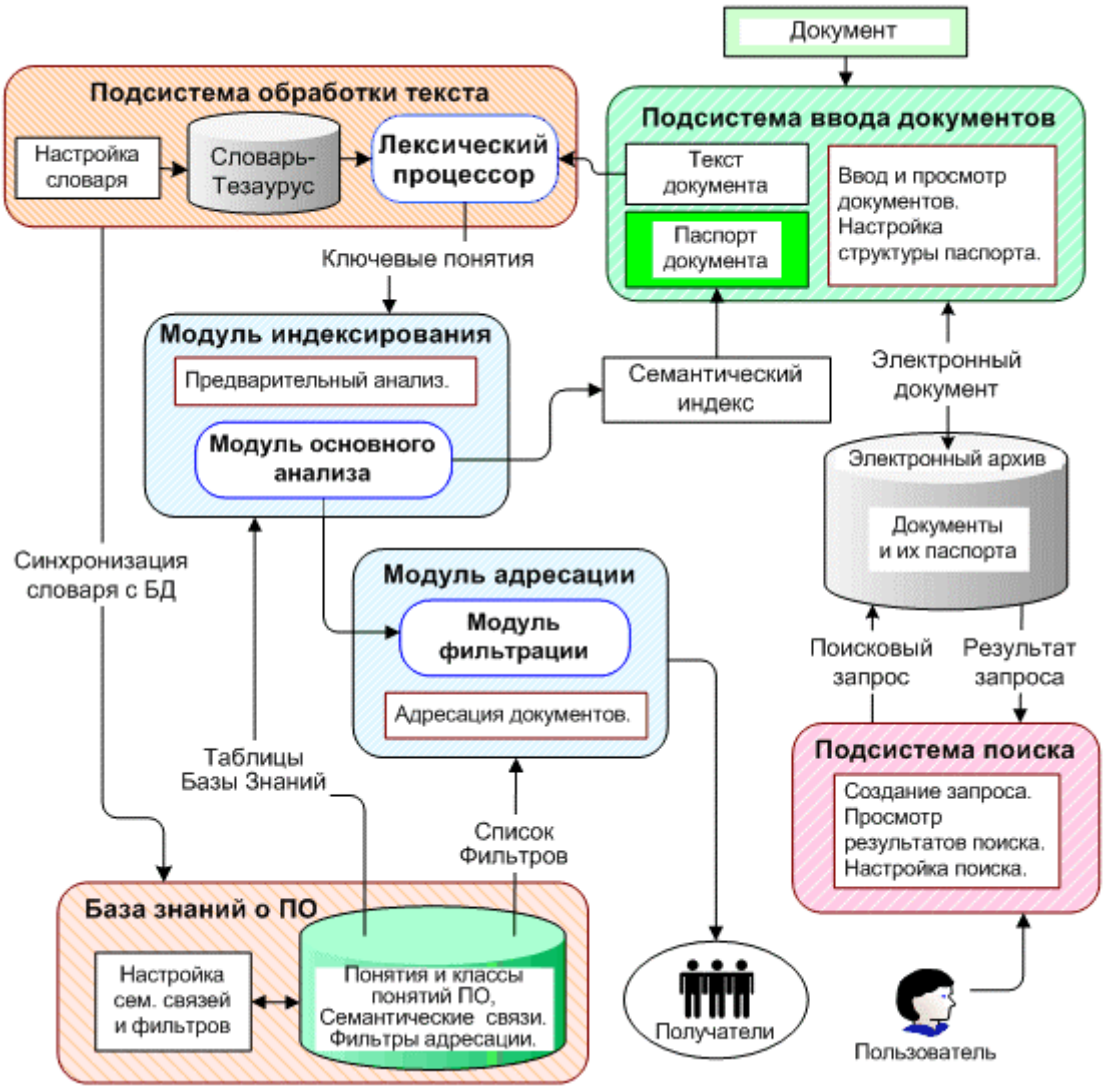

Рис.1. Архитектура и потоки данных в системе InDoc.

Знания о предметной области (ПО) и языке документов вносятся в базу знаний через автоматизированные рабочие места, которые позволяют эксперту корректировать и пополнять базу знаний.

В процессе автоматической обработки текст документа поступает на вход лексического процессора, который выделяет из текста ключевые понятия ПО и передает их модулю индексирования. Модуль индексирования осуществляет основной анализ документа, результатом которого является совокупность фактов, отражающих основное содержание (тему) документа.

Для обеспечения рассылки документов их фактическим получателям разработан модуль адресации, который фильтрует полученные в результате основного анализа факты, сравнивая их с ассоциированными с пользователями фильтрами.

По завершении этапа автоматической обработки электронный документ направляется адресатам и в электронный архив. Получение необходимых пользователю документов из электронного архива обеспечивается подсистемой поиска.

2. Представление знаний в системе

Технология семантического индексирования опирается на знания о подъязыке документов и знания о структуре и деятельности организации.

2.1. Знания о предметной области. В основе базы знаний лежит модель предметной области, которая отражает систему основных понятий и семантические отношения между ними. Иерархии классов понятий и заданные на них

семантические отношения позволяют представить структуру высказывания из предметной области в виде Факта, множество которых составляет пропозициональное содержание документа.

2.1.1. Иерархия понятий. При создании иерархии понятий ПО принимаются во внимание следующие соображения: семантическая роль понятия в структуре высказывания; тезаурусные отношения между понятиями – их родовидовая общность и отношение структурной вложенности.

Иерархия Работ включает основные классы, представляющие определенные пользователем виды деятельности (например, Подготовка производства, Проектирование, Строительство).

Иерархия Объектов, используемых или создаваемых в процессе работы, включает Ресурсы (природные, материально-технические), Документы (проектная, проектно-сметная, техническая документация и т.п.) и Объекты строительства.

Иерархия Объектов строительства выделяется ввиду особой семантической роли отнесенных к ней понятий – они представляют локализацию Работ: стройки, участки строек, подобъекты строек. В этой иерархии понятия сгруппированы в классы с учетом структуры и способа идентификации соответствующих наименований. Так, класс Именованных участков (*участок Запад–2*) противопоставлен классу Параметризованных участков, которые идентифицируются линейным параметром (*участок км 0 – 1003*).

Иерархия Организаций содержит классы Заказчик, Инвестор и Партнеры. Заказчик – это организация-пользователь системы, а партнеры – множество организаций, взаимодействующих с Заказчиком по различным аспектам деятельности и объединенных в подклассы в соответствии с их функциями (Проектировщики, Строители, Субзаказчики, Поставщики).

2.1.2. Семантические отношения. Знания о сочетаемости понятий представлены в виде отношений, задающих ограничения на допустимые сочетания понятий указанных классов и их значений в Фактах.

Отношение вложенности – отношение "часть–целое", которое задается на множестве конкретных понятий из иерархии Объектов строительства и позволяет идентифицировать в тексте объекты сложной структуры:

участок км 0–1003 <ПарамУчасток> первой очереди <Пусковой комплекс> газопровода "Север–Европа" <Стройка>

Объектное отношение – связь <Работа – Объект>, которое задается на классах понятий и формирует семантическое ядро Факта (Функцию), однозначно определяющую Вид деятельности:

перевод лесных земель <ПрирРесурс> *в нелесные* <Природопользование> => ПодготовкаПроизводства

Агентивное отношение – связь <Организация – Функция>. Это отношение характеризует различные классы иерархии Организаций как субъектов деятельности.

Поставщик – Комплектация, Строитель – Строительство, Субзаказчик – Комплектация, Субзаказчик – Строительство

2.2. Лингвистические знания. Знания, необходимые для лингвистической обработки текста документа, представлены в словаре-тезаурусе в виде библиотеки именованных шаблонов, настроенных на предметную область и жанр документов.

Наиболее многочисленную группу в словаре составляют ***предметные шаблоны***. Иерархия классов предметных шаблонов соответствует иерархии классов ПО базы знаний и покрывает все слова и словосочетания, которые выражают необходимые при анализе документа ключевые понятия. Каждый шаблон объединяет в себе множество синонимических выражений одного понятия. Имя шаблона соответствует нормализованному виду слова или словосочетания, представляющего данное понятие.

Жанровые шаблоны используются для жанровой сегментации документа, которая является необходимой составляющей автоматической обработки, поскольку позволяет выделить в документе функциональные информационные блоки (Отправитель, Основной текст). Процесс анализа использует границы Основного текста для извлечения ключевых понятий, на базе которых анализируется пропозициональное содержание текста документа и выявляется его тема.

3. Семантическое индексирование

Семантический анализ состоит в установлении семантических отношений между составляющими высказываний в Основном тексте, что позволяет представить содержание документа в виде совокупности упомянутых в нем Фактов (подробное описание процедуры см. в [1]).

3.1. Содержание семантического индекса. **Основной вопрос, связанный с семантическим индексированием, – какая информация должна быть представлена в семантическом индексе, чтобы он отражал тематику документа и был удобен для просмотра.**

Ключевые понятия текста документа представлены в структуре Факта в их взаимосвязи, что позволяет более точно определить тему документа. Так, чтобы правильно идентифицировать тему 'Проектирование', необходимо убедиться, что в составе высказываний текста имеются определяющие данный вид деятельности функциональные связи на основе объектных отношений: *проектные работы по ГИС 1.1*, *разработка проекта* (для сравнения, функциональная связь *выслать проектные материалы* не соответствует никакому из основных видов деятельности)

Тема релевантного множества документов может быть сформулирована пользователем как отдельное понятие вне ситуации (например, вся информация, касающаяся конкретного объекта строительства *ГИС 1.1 газопровода Запад–Восток*), либо в виде описания конкретной или обобщенной ситуации (*экологическая экспертиза проекта по ГИС 1.1 газопровода Запад–Восток* vs. *проектирование объектов газопровода Запад–Восток*). Для корпуса документов архива после согласования с пользователем был определен следующий набор полей семантического индекса:

- список ключевых понятий;
- тема документа;
- организация-отправитель;
- виды деятельности;
- объекты строительства.

В поле <Тема> содержание документа представляется в виде множества Фактов, с большей или меньшей степенью конкретности описывающих основное содержание документа. Дополнительные три поля представляют прагматически важные для пользователя аспекты темы документа.

3.2. Определение Видов деятельности. Определение Видов деятельности, о которых идет речь в документе, – основной этап в процессе тематического анализа. Исходя из информации базы знаний о возможных объектных связях между ключевыми понятиями классов Работа и Объект, строятся Функции в границах формальных сегментов (предложений). Вид деятельности представляет собой один из базовых классов Работ, и рассматривается как тематическое обобщение семантической связи <Работа – Объект>.

экологическая экспертиза <Экспертиза> *проекта* <ПроектДокумент> => Проектирование

доработка <УниверсРабота> *электростанции* <ЭлектроОборудование> => Комплектация

Функции, которые не определяют никакого Вида деятельности (*выслать проектную документацию*), далее не рассматриваются, как не относящиеся к теме.

3.3. Идентификация Объектов строительства. При определении значений поля <Объект строительства> семантический анализ решает задачу идентификации выделенных в тексте документа объектов строительства с объектами, представленными в иерархии вложенности в базе знаний.

Параметризованные и именованные объекты не всегда воспроизводятся авторами документов в точном соответствии с номенклатурой объектов строительства, зафиксированной в базе знаний. В таких ситуациях требуется уточнение указанных приблизительных значений линейных параметров участка или определение эталонного имени по указанным линейным параметрам (как в примере, приведенном в п.3.4, где *участок Запад-1* идентифицирован в тексте линейным отрезком *участок км 1286 - км 1310*).

Определение Объекта строительства требует восстановления иерархии вложенности объектов документа путем сравнения с эталонной иерархией базы знаний. Каждая пара ключевых понятий, отнесенных в тезаурусе к иерархии Объектов строительства и удовлетворяющая определенным требованиям порядка слов, проверяется на предмет наличия между ними отношения вложенности (с учетом транзитивности). Это позволяет собрать Объекты сложной структуры, представленные линейными цепочками наименований, совокупность которых образует дерево (множество деревьев) объектов строительства данного документа. Результирующими являются те объекты базы знаний, которые соответствуют листьям полученных древесных структур.

3.4. Тема документа. При анализе корпуса деловых писем было выявлено, что основное содержание документа определяется деятельностью Партнера-Отправителя. Это позволяет рассматривать Факты, ориентированные на Отправителя, как основные, относящиеся к Теме документа. Исходя из этого, дальнейшая процедура состоит в проверке сочетаемости Видов деятельности с классом Организации-Отправителя и установлении агентивных связей. В результате определяется окончательное множество значений поля <Вид деятельности>. Объединение

полученных структур с идентифицированными Объектами строительства дает множество основных Фактов документа. Разбиение этого множества по Видам деятельности позволяет представить Тему как совокупность Тематических Фактов:

Для прохождения газопровода Запад-Восток, км 1286 - км 1310 в скальных грунтах требуется применение скального диета в количестве 210 листов стоимостью 220 тыс.руб. взамен резино-тканевых листов.

Тематический Факт 1

Вид деятельности: Строительство

Работа: СтроительствоМГ[прохождение] – Объект: Стройка[газопровод Запад-Восток]

Работа: СтроительствоМГ[прохождение] – Объект: ПарамУчасток[участок Запад–1]

Работа: УниверсРабота[использование] – Объект: Материал[скальный диет]

Работа: УниверсРабота[замена] – Объект: Материал[резино–тканевый лист]

Организация–Отправитель: Субзаказчик[Севергазпром]

Объект строительства: [газопровод Запад-Восток]/ [участок Запад–1]

В результате процедуры обработки и генерации семантического индекса завершается оформление электронного паспорта, и документ вместе со своим паспортом отправляется в электронный архив.

4. Поиск документов в электронном архиве

Подсистема поиска предоставляет пользователю возможность обратиться в электронный архив организации за необходимой справочной информацией.

4.1. Поисковый запрос. Принято считать, что для пользователя наиболее естественно и удобно формулировать свои запросы на естественном языке. Однако мнение о максимальном удобстве естественно-языкового доступа к архиву можно оспорить, если речь идет об организации архива в рамках корпоративной информационной системы, база знаний которой содержит огромные объемы такой достаточно сложно организованной концептуальной информации, как иерархия объектов строительства или номенклатура материально-технических ресурсов. Исходя из этого, в системе InDoc предусмотрены формализованные поисковые запросы, в которых ограничения на значения полей индекса задаются в терминах структур базы знаний с использованием иерархических интерфейсных форм.

Структура формализованного запроса соответствует структуре индекса и содержит в себе шаблон Факта. Интересующие пользователя поля запроса заполняются именами понятий или классов базы знаний. Документ удовлетворяет поисковому запросу, если в семантическом индексе документа присутствует Факт, удовлетворяющий шаблону запроса.

Кроме того, пользователю предоставляются средства для создания и сохранения так называемых предопределенных запросов. Предопределенный запрос служит для выражения достаточно типичной и регулярно возникающей информационной потребности, например:

«В каком состоянии находилась комплектация материально–техническими ресурсами объекта строительства X в период Y?».

В предопределенном запросе выделяются две части: фиксированная и изменяемая. Фиксированная часть заполняется при создании предопределенного запроса, а изменяемая — непосредственно перед запуском запроса на исполнение. Так, в приведенном примере:

Фиксированная часть: Вид деятельности = «комплектация» & Тип документа = «деловое письмо»
Переменная часть: Объект строительства = <…> & Дата поступления = диапазон <…>

Поисковый запрос пользователя преобразуется в SQL-запрос, который возвращает список удовлетворяющих запросу электронных документов архива.

4.2. Поиск и просмотр документов. При поиске запрос сопоставляется с содержанием индекса документов архива. Процедура сравнения учитывает наследование в иерархии классов. Если в составе запроса присутствует альтернативная спецификация значений для некоторой составляющей, то достаточно найти соответствие хотя бы для одной из возможных комбинаций.

Документ удовлетворяет поисковому запросу, если для каждого непустого поля запроса нашлось хотя бы одно значение из аналогичного поля индекса. В частности, это значит, что в Теме индекса обнаружен соответствующий запросу Тематический Факт.

Пользователю предоставляются электронные паспорта документов, индексы которых успешно сопоставились с формальным представлением запроса. Для оценки результатов поиска существует возможность просмотра электронных паспортов найденных документов и сортировка документов по значениям полей. Допускается фокусировка поискового запроса (уточнение по иерархии понятий, сужение диапазона дат и т.п.) — в этом случае осуществляется дополнительный поиск в массиве документов, найденных на предыдущем шаге. Если результат удовлетворяет пользователя, осуществляется вывод найденных документов.

5. Поддержка актуальности индекса

В последнее время в области информационного поиска рассматриваются задачи обработки динамических массивов документов и связанные с ними проблемы актуальности используемой системой информации ([5]).

При создании системы обработки и архивирования потока входящих документов возникает проблема поддержки актуальности индекса архива. Поскольку система InDoc основана на знаниях, необходимо предусмотреть возможность как чисто терминологических изменений, так и модификаций в знаниях о ПО. С течением времени появляются расхождения между текущим состоянием базы знаний и ранее построенным индексом. В такой ситуации поисковый запрос, который генерируется в рамках текущего состояния знаний, может неправильно работать с "устаревшим" индексом. Для восстановления работоспособности индекса используется механизм *локальной переиндексации,* т.е. частичной переиндексации небольшой части архива.

Выделяются три категории изменений в базе знаний, которые требуют различных подходов при решении проблемы актуальности индекса.

5.1. Преобразование системы знаний. К этой группе относятся глобальные изменения в системе знаний, которые, соответственно, требуют глобальной переиндексации всего архива или большей его части (если такие кардинальные изменения вообще допустимы в рамках единого архива):

- существенные изменения иерархии классов (перенос ветви иерархии, добавление или удаление базовых классов ПО);
- переопределение семантической сочетаемости классов понятий;
- перестройка иерархии вложенности объектов строительства.

Такая перестройка базы знаний требует ***глобальной переиндексации***, что является нештатной ситуацией и осуществляется разработчиками совместно с экспертом в режиме технического обслуживания системы.

5.2. Корректировка системы знаний. К этой группе относятся локальные изменения с целью коррекции некоторого фрагмента базы знаний (например, поправки, вносимые экспертом в процессе опытной эксплуатации системы):

- удаление периферических классов понятий, которое сопровождается удалением ограничений сочетаемости этих классов;
- удаление сочетания классов понятий;
- модификация значения вида деятельности в функциональном сочетании;
- удаление конкретных понятий и соответствующих им терминов;
- удаление объекта строительства или изменение параметров объекта строительства.

При таких изменениях применяется ***автоматическая локальная переиндексация*** архива. После корректировки базы знаний автоматически запускается процедура диагностики индекса, результат которой отражает наличие/отсутствие и тип расхождений между содержанием индекса и текущим состоянием базы знаний. В зависимости от типа обнаруженных несоответствий, для каждого документа определяются те части индекса, которые нужно заново индексировать. Например, при удалении некоторого понятия из словаря, достаточно удалить это понятие и включающие его структуры из индекса документа.

5.3. Пополнение системы знаний. К этой категории отнесены изменения состояния знаний, которые возникают с течением времени и должны учитываться при индексировании вновь поступающих на обработку документов:

- пополнение существующей иерархии понятий новыми классами, возможно, с необходимостью фиксации дополнительных ограничений сочетаемости этих классов;
- появление новых понятий и соответствующих им терминов (например, использование новых видов оборудования или введение новой организации в число деловых партнеров);
- пополнение иерархии вложенности объектов строительства;
- сдвиги в соотношении "понятие – языковое выражение"; отмечено два вида терминологических сдвигов:
- диахроническая синонимия терминов – использование нового языкового выражения для существующего понятия, в результате, по крайней мере, некоторое время два термина существуют параллельно (*ОАО Газпром* и *РАО Газпром* для понятия 'Газпром');
- диахроническая омонимия термина, например, аббревиатура *ПК* для ключевого понятия 'пикет' с некоторого момента времени стала использоваться и для понятия 'пусковой комплекс'.

Поскольку изменения третьего типа вносят в базу знаний новые элементы, семантический индекс документов архива этих элементов не содержит. Отсюда можно сделать вывод, что никакой переиндексации при пополнении базы знаний не требуется. Однако новые элементы знаний могут функционировать в документах и попадать в архив с некоторым опережением по отношению ко времени модификации базы знаний. В связи с этим предусматривается возможность ***ручного запуска локальной переиндексации*** по специальному указанию администратора системы: обновление семантического индекса документов, поступивших за последний небольшой промежуток времени.

6. Развитие системы InDoc

Основное направление развития системы InDoc связано с повышением ее эффективности. Рассматривается организация электронного архива в виде единого информационного пространства, которое содержит как данные, извлеченные из документов архива на этапе семантического анализа, так и постоянно пополняющуюся базу знаний, отражающую динамику предметной области в связи с развитием деятельности и изменениями в структуре организации.

Для решения этой задачи предполагается создавать некоторые базовые объектные структуры, отражающие состояние дел на каждом объекте строительства и хранить в них информацию, на основе которой документ в процессе индексирования получает связь с данным объектом строительства, партнером и типом работы (видом деятельности), упомянутыми в документе. Пример структуры объекта:

Объект строительства:

список вложенных ***объектов строительства***;

список выполняемых ***работ***:

состояние работы;

список ***организаций,*** выполняющих работу;

список ***документов*** по выполняемой работе;

список используемых ***материально–технических ресурсов***.

Построенное таким образом информационное пространство обеспечит эргономичную навигацию по всему архиву и эффективный поиск в нем необходимых документов.

Список литературы

1. Кононенко И.С., Сидорова Е.А. Обработка делового письма в системе документооборота //Труды международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям, т.2 - Протвино, 2002 – с 299-310.
2. Salton, G. Automatic Text Processing: The transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley Publishing Company, Inc., 1989.
3. Лукашевич Н.В., Добров Б.В. Тезаурус для автоматического концептуального индексирования как особый вид лингвистического ресурса //Труды международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям, т.2 - Аксаково, 2001 – с 273-279.

4. Basili R., Pazienza M.T., and Mazzucchelli L. An Adaptive and Distributed Framework for Advanced IR // Content-Based Multimedia Information Access. RIAO'2000 Conference Proceedings, v.2, 2000, – pp. 908–922.

5. Wolinski F., Vichot F., and Stricker M. Using Learning–based Filters to Detect Rule–based Filtering Obsolescence // Content-Based Multimedia Information Access. RIAO'2000 Conference Proceedings, v.2, 2000, – pp. 1208–1220.