

СИСТЕМА ПОИСКА И КЛАССИФИКАЦИИ ДОКУМЕНТОВ С ИСПОЛЬЗОВАНИЕМ ИЕРАРХИЧЕСКОГО СЛОВАРЯ ТЕМ

А. Ф. Гельбух, Г. О. Сидоров, А. Гузман-Аренас

Лаборатория естественного языка Центра Компьютерных Исследований (CIC)
Национального Политехнического Института (IPN) Мексики
LabLN, CIC, IPN, Av. Juan Dios Batiz s/n esq. Mendizabal, Zacatenco 07738, DF, Mexico.
{gelbukh, sidorov, aguzman}@pollux.cic.ipn.mx.

ABSTRACT

A statistical method of document classification driven by a hierarchical topic dictionary is proposed. The method uses a dictionary with a simple structure and is insensitive to inaccuracies in the dictionary; the dictionary is easily trainable on a manually classified document collection and even automatically translatable into different languages. A common sense-complaint way of assignment of the weights to the topics is presented; these weights are taken into account when the system answers the question “what are the topics of the given document?” and neither too narrow nor too broad answer is desirable. A system Classifier based on the discussed method is described.

ВВЕДЕНИЕ

В данной работе* мы рассматриваем задачу поиска и классификации документов (текстов) по их главным темам: например, одни документы могут быть про *животных*, а другие про *промышленность*. Данная задача существенна при информационном поиске, классификации потоков документов (например, входящих документов в крупном учреждении), фильтрации потоков документов, например, новостей в Интернет, и во многих других приложениях.

Рассматривается случай, когда набор основных тем фиксирован заранее, хотя его объем может быть большим. Таким образом, мы не рассматриваем задачи автоматического обучения, извлечения ключевых тем из тела самого документа и т.д. Наша задача сводится к отнесению документа к одной (или нескольким) из тем, заданных в словаре системы. Результатом является мера (в процентах) соответствия документу каждой из имеющихся в словаре тем.

Естественно, в качестве основной темы выбирается та, которая получила наилучшую оценку. В задаче собственно классификации (скажем, направление входящего документа в один из отделов учреждения) документ относится именно к этой категории. Однако для наших целей мы предпочитаем рассматривать задачу

* Работа выполнена при частичной поддержке REDII-CONACyT, CONACyT, грант 26424-A, DEPI-IPN и SNI, Мексика.

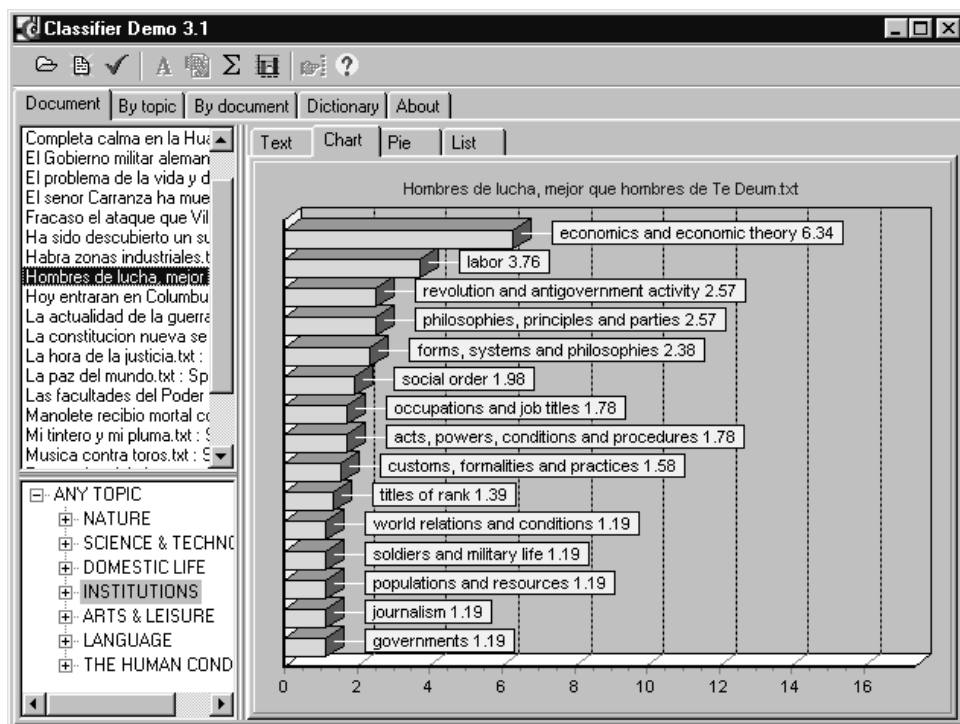


Рис. 1. Гистограмма тем для документа.

«нечеткой» классификации, при которой документ относится сразу к нескольким категориям, с соответствующими весами, — то есть характеризуется вектором, или гистограммой, весов тем.

На Рис. 1 показано окно разработанной нами программы Classifier с гистограммой тем одного из документов (выбранного в списке в левой части окна). В настоящее время программа Classifier снабжена словарями для испанского, английского и французского языков; на наших иллюстрациях показаны документы на испанском языке. В то же время названия тем рассматриваются нами как имена, не зависящие от рабочего языка документа и обозначаемые на иллюстрациях английскими словами.

При такой классификации возникает проблема выбора разумной степени детализации темы. Например, при классификации интернетовских новостей для «среднего» пользователя такие темы, как *животные* и *промышленность* представляются вполне уместными, в то время как при классификации статей по зоологии хищников такой способ классификации приведет к выдаче бесполезного ответа о том, что все статьи относятся к одной и той же категории *животные*. С другой стороны, возвращаясь к классификации интернетовских новостей, наш «средний» пользователь очень удивился бы, увидев, что система отнесла статью про лошадей к разряду *непарнокопытные*.

В статье мы кратко рассмотрим структуру словаря, используемого при построении гистограммы тем, способы решения вышеупомянутой проблемы оптимального выбора детализации темы и некоторые применения изложенных методов, а также некоторые практические аспекты составления словаря, используемого программой.

ИЕРАРХИЧЕСКИЙ СЛОВАРЬ ТЕМ

Идея использования иерархии тем была предложена, например, в [1]. В том виде, в каком мы приводим ее в данной работе, она была предложена нами в [2, 3].

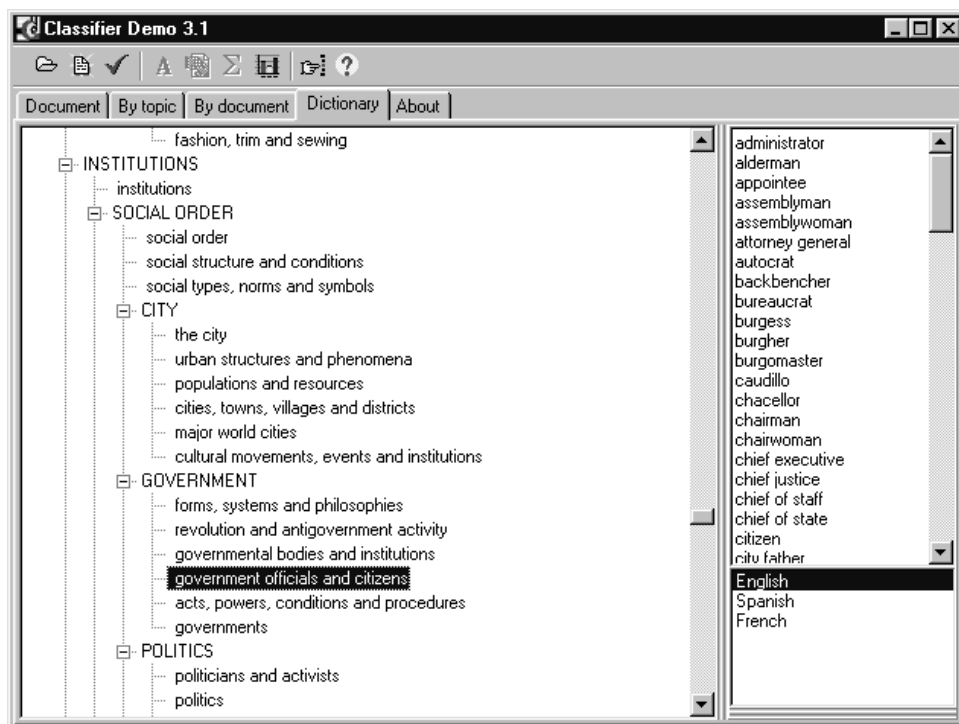


Рис. 2. Иерархический словарь системы.

Технически словарь разделен на две части: квазисинонимические группы, формирующие элементарные темы, или концепты, и иерархия таких концептов. Формально же каждое слово вполне может рассматриваться как отдельный узел общей иерархии; в этом случае иерархия задается непосредственно на словах, а не на квазисинонимических группах.

Квазисинонимическую группу составляют слова, относящиеся к ситуации, которую описывает тема. Например, тема *религия* включает такие слова, как *церковь*, *свеча*, *пон*, *Библия*, *молитва*, *тигрим* и т.д. На Рис. 2 показано окно программы с фрагментом словаря. Технически наша программа трактует словосочетания так же, как и отдельные слова. Так, терминологическое словосочетание *линейное программирование* может входить в группу *исследование операций*, в которую отдельные слова *линейный* и *программирование* не входят.

Поскольку в нашем словаре не требуется специфицировать типы связей между узлами иерархии, составление такого словаря является в известном смысле более простой задачей, чем составление традиционной семантической сети, такой как [4]. Так, составление большого такого словаря можно доверить лицам без специальных лексикографических навыков (студентам); более того, есть надежда на автоматическое пополнение такого словаря на основе корпуса текстов.

Необходимо заметить, что словарь рассматриваемого типа не есть просто урезанный вариант семантической сети вроде WordNet или FACTOTUM [4] с удаленными именами связей, поскольку в нашем словаре могут оказаться связанными слова, не имеющие непосредственной семантической связи в традиционных сетях.

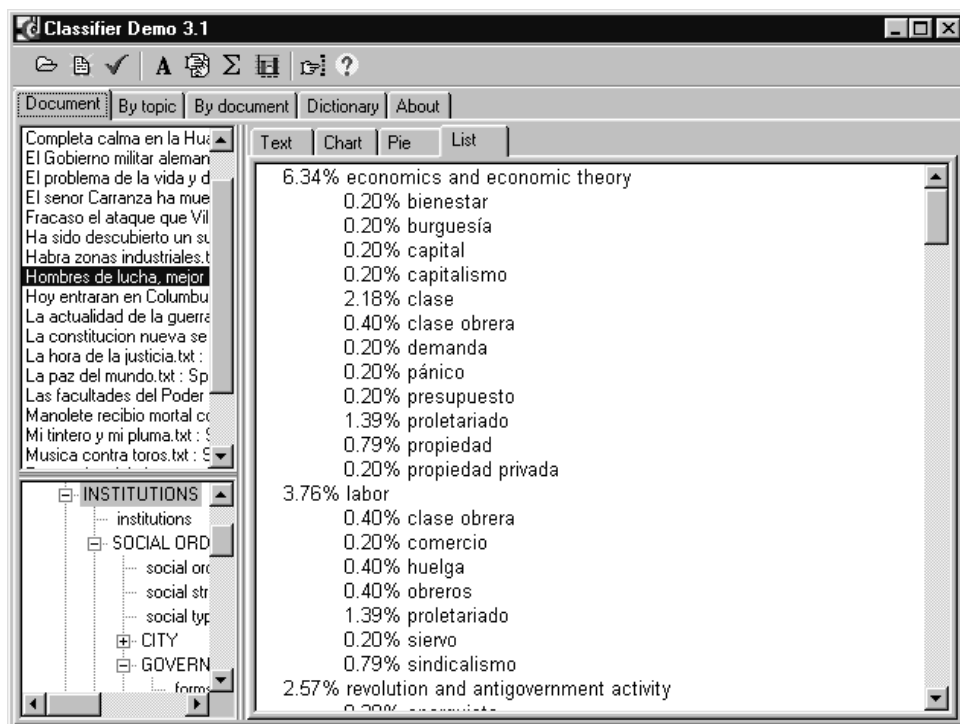


Рис. 3. Число вхождений ключевых слов в документ.

ОСНОВНОЙ АЛГОРИТМ КЛАССИФИКАЦИИ И ВЕСА СВЯЗЕЙ

Алгоритм классификации с применением рассмотренного словаря состоит, как и сам словарь, из двух частей: определения весов терминальных узлов и продвижения весов вверх по иерархии.

Первая часть алгоритма отвечает для каждой темы на вопрос: в какой степени данный документ релевантен для данной темы? В нашей нынешней реализации степенью релевантности, или весом, группы слов является доля в данном документе слов, входящих в эту группу, то есть число вхождений всех слов данной группы в документ, поделенное на число слов в документе. Веса тем не нормализуются, поскольку один и тот же документ может быть посвящен одновременно нескольким темам, то есть суммарный вес тем для одного документа может быть и больше единицы.

Однако простой подсчет числа вхождений слов не является единственно возможным методом. Так, в будущем мы не исключаем возможности включения в словарь активных, то есть процедурных, элементов. Например, для определения того, что текст является официальным документом – скажем, заявлением или финансовой ведомостью, может быть необходимым проанализировать формат документа, его заголовок, наличие определенных полей.

Вторая часть алгоритма отвечает за продвижение весов вверх по иерархии. Так, если документ говорит о *кошках* и *собаках*, то он также говорит о *животных* и о *природе*. Ниже мы рассмотрим способы подсчета весов таких нетерминальных тем.

Ясно, что для достижения разумных результатов при определении релевантности тем для документа необходимо использовать определенные числовые коэффициенты, или веса. Мы применяем два типа таких коэффициентов: веса связи отдельных слов (или отдельных подтем) с темами и веса важности тем, которые мы называем различительными весами по причинам, объясняемым ниже. Первый тип весов

приписывается паре узлов в иерархии (или паре, состоящей из слова и группы), а второй – отдельным узлам.

Назначение этих весов – дать количественную характеристику того, насколько именно включенное в группу слово «голосует» за свою группу, или насколько подтема «голосует» за тему, которой она подчинена в иерархии — то есть насколько сильно связаны слово и тема или тема и подтема. Например, слово *свеча* меньше относится к теме *религия*, чем слово *Библия*. Другими словами, если текст содержит слово *Библия*, то он скорее всего о религии, а если слово *свеча* – то он с меньшей вероятностью окажется о религии. Следовательно, связи между *свечой* и *религией* может быть поставлен в соответствие определенный коэффициент. Поскольку эти коэффициенты привязаны к дугам иерархии, они снабжаются двумя индексами: вес w_k^j соответствует связи между узлом k и вышестоящим узлом j , за который узел k «голосует».

Использование этих весов в алгоритме очевидно. Веса связи w_k^j слов с темами используются при определении веса r^j темы j на основе суммы весов r^k входящих в нее слов (или подтем) k по следующей формуле:

$$r^j = \sum_{k \in j} w_k^j r^k, \quad (1)$$

(здесь мы допускаем некоторую вольность обозначений), то есть $религия = 0.9 \times Библия + 0.3 \times свеча + \dots$. Подсчет весов начинается с терминальных узлов иерархии — отдельных слов, для них вес r^j есть просто число вхождений данного слова j в документ, далее веса продвигаются вверх по иерархии.

Как определить веса w_k^j ? В [7] они расставляются вручную на основании интуиции человеком-экспертом на стадии подготовки словаря. Выбор веса, по-видимому, основан на сравнении частоты слова в документах на данную тему с частотой его появления в самых «общих» документах для «любой» темы, например, газетной смеси. Однако такой способ обладает рядом недостатков, среди которых — большой объем работы, отсутствие ясной методики назначения весов, склонность экспертов недооценивать омонимичность терминов (а термины очень часто омонимичны, ср. *корень* — слова, степени, растения, проблемы, *программа* — компьютерная, экономическая, заседаний).

В данной работе мы приведем способ их автоматического вычисления. Сначала рассмотрим случай, когда мы располагаем набором текстов, снабженных (скажем, вручную) величинами релевантности отдельных тем (впрочем, они могут быть и просто 0 — нерелевантна и 1 — полностью релевантна). Тогда коэффициент связи слова с темой может быть определен как средняя релевантность данной темы для документа, содержащего одно или несколько вхождений данного слова:

$$w_k^j = \frac{\sum_{i \in D} r_i^j n_i^k}{\sum_{i \in D} n_i^k} \quad (2)$$

где суммирование проводится по всем документам i набора D , r_i^j — релевантность документа i для темы j , а n_i^k — число вхождений слова (или вес подтемы) k в документе i . Данная формула отвечает на вопрос: какова средняя релевантность темы *религия* для документов, содержащих слово *свеча*? Доказательство данной формулы основано на применении формулы Байеса к условной вероятности того, что документ, содержащий данное слово, релевантен для данной темы.

Возможен и компромисс между двумя методами — ручным и автоматическим. В случае, если отдельный корпус отобранных по каждой теме документов отсутствует, можно использовать составленный экспертом словарь предметной области — набор слов, которые он считает релевантными для заданной темы — и назначить им веса автоматически. Такой метод применим только при небольшом числе тем в словаре, причем тем достаточно узких. В этом случае есть основания предположить, что тексты с данной темой вообще не встречаются в контрольном корпусе (газетной смеси). Тогда, имея в виду, что слово оказалось включенным в словарь, и, значит, по крайней мере имеется один документ, соотнесенный с данной темой, можно упростить выражение для оценки весов следующим образом:

$$w_k^j = \frac{1}{\sum_{i \in D} n_i^k} \quad (3)$$

поскольку можно предположить, что знаменатель в (2) для очень узких тем равен 1: имеется только один релевантный документ. Естественно, это дает вес слова, «голосующего» за определенную тему, тем меньше, чем больше его частота в газетной смеси. Например, служебные слова вроде *и* или *а* имеют практически нулевой вес для каждой темы, а слово *карбюратор* достаточно большой вес в любой теме, к которой оно вообще отнесено экспертом.

РАЗЛИЧАЮЩИЕ ВЕСА

Описанный выше алгоритм классификации (1) может применяться для ответа на вопрос «Является ли данный документ документом про *животных*». Однако если попытаться поставить вопрос по-другому: «Про что этот документ?», то при буквальном применении формулы (1) ответ будет всегда одинаков: про *объекты* и *действия* (самые верхние уровни иерархии). В то время как осмысленным был бы ответ, например: о *ракообразных*, или о *животных*, или о *живых существах*, или о *природе* — в зависимости от ситуации.

Как определить, в чем заключается эта «ситуация»? Вероятно, это — степень детализированности, с которой пользователь «видит» предметную область. Для биолога лучшим будет ответ *ракообразные*, а для рядового читателя газеты — *природа*. Как можно догадаться, с кем мы имеем дело, не спрашивая этого в явном виде? Тем более, что такой вопрос сложно сформулировать, и как раз самому массовому пользователю он вообще вряд ли будет понятен.

Наша гипотеза состоит в том, что «универсум» пользователя — это сама база документов, в которой он собирается что-то искать или классифицировать, т.е. что пользователь является специалистом как раз по его собственной базе. Таким образом, веса тем в нашей системе зависят от текущей базы.

Основное требование к весам — их *различительная сила*: тема должна соответствовать разумному подмножеству документов, не слишком большому и не слишком маленькому. Тема, которая соответствует всем документам базы данных, не несет никакой информации. Таким образом, вес w^j узла j может быть вычислен как вариация релевантностей r_i^j данной темы по всем документам базы. В текущей системе мы используем следующее выражение:

$$w^j = \sum_{i \in D} (r_i^j - M)^2, \quad M = \frac{\sum_{i \in D} r_i^j}{|D|} \quad (5)$$

здесь M - среднее значение r_i^j для текущей базы D , и r_i^j определяется по описанному выше алгоритму, не принимая во внимание значение w^j . Вероятно, можно составить и формулу, более обоснованную с точки зрения теории информации.

При таком подходе, скажем, в базе данных про биологию вес тем *животные*, *живые существа*, *природа* низок, потому что практически все документы в одинаковой степени упоминают эти темы. С другой стороны, в «газетной смеси» вес этих тем достаточно высок, а, наоборот, вес темы *ракообразные* низок. Следовательно, одна и та же статья про особенности планктона, обитающего в Карибском море будет классифицирована как относящаяся к *природе* и *географии* в базе газетных сообщений, и как относящаяся к *ракообразным* в базе по биологии.

ПРАКТИЧЕСКИЕ ВОПРОСЫ РЕАЛИЗАЦИИ

Описанный выше подход послужил основой для реализации в нашей системе CLASSIFIER нескольких полезных функций.

Система может определить основные темы документа, см. Рис. 1. Это соответствует задаче классификации. Также система позволяет сгруппировать документы по темам, тем самым отвечая на вопрос: какие документы наиболее релевантны для заданной темы? Это приближенно соответствует задаче информационного поиска.

Интересное приложение метода — классификация документов по сходству с учетом конкретного «аспекта» — темы. Очевидно, что документ, в котором говорится об использовании животных в военных целях, и документ про то, как правильно кормить животных, похожи с точки зрения биолога (в них говорится про *животных*), но различны с точки зрения военного. Сравнение производится на основе сравнения весов всех подтем данной темы в двух документах.

Самая трудная проблема при реализации описанного метода состоит в компиляции словаря. Применяемый в нашей системе иерархический словарь (для английского языка) мы создали на основе различных источников, включая тезаурус Роже, WordNet, FACTOTUM SemNet и некоторых других словарей.

Однако нам были нужны еще словари для французского и испанского языков. Оказалось, что наш статистический метод настолько устойчив к наличию «шума» и ошибок в словаре, что даже автоматический перевод английского словаря дает очень хорошие результаты. Простейшая процедура перевода без учета слов контекста (т. е. не различающая значения) позволила получить пригодные для использования словари. Дальнейшее улучшение французского и испанского словарей достигается при помощи контекстного перевода, описанного в [5, стр. 363].

РЕЗУЛЬТАТЫ И НАПРАВЛЕНИЯ ДАЛЬНЕЙШИХ ИССЛЕДОВАНИЙ

В целом результаты, полученные нами экспериментально, хорошо согласуются с оценками экспертов. Однако, мы столкнулись и с некоторыми проблемами при реализации метода. Большинство из них, как и почти все проблемы обработки текста, связано с неоднозначностью.

Бывает, что частотное слово, взятое без контекста, является важным для какой-либо темы: английское существительное *well* (*скважина*) является важным термином *нефтедобычи*, существительное *do* (*стиль прически*) - термин *парикмахерского дела*, существительное *in* (*влияние*) — *политики*, и др. Однако, выражение (2) или (3)

присваивает слишком малый вес таким ключевым словам. Для решения этой проблемы мы планируем добавить в нашу систему маркер частей речи. Для более подробного анализа возможно подключение разработанного нами синтаксического анализатора [6], однако это существенно замедлит работу системы.

И, наконец, мы планируем попробовать итерационную процедуру определения весов связей. Действительно, формулы (1) и (2) являются взаимно зависимыми. В некоторых случаях доступная информация сводится к определенным человеком-экспертом весам релевантности тем для документов, и тогда по формуле (2) могут быть определены веса связей в иерархии. В других случаях изначально вручную назначаются именно эти веса. В большинстве же случаев может быть получена и та, и другая информация, однако в очень небольших объемах. Возможно, итеративная переоценка этих двух параметров — r^j и w_k^j — путем последовательного применения формул (1) и (2) может привести к положительным результатам. Необходимы эксперименты, чтобы выяснить, сойдется ли итеративный процесс к разумному результату или к некоторой точке, вообще не зависящей от начального распределения весов — в этом случае итеративный метод неприменим [8].

Хотя точность работы алгоритма все еще ниже желаемой, его результаты хорошо совпали с экспертной оценкой, данной человеком. Метод является практически применимым в том смысле, что он достаточно нечувствителен к неточностям в словаре, а также в том смысле, что структура словаря настолько проста, что словарь может быть полуавтоматически извлечен из корпуса размеченных по темам документов и даже автоматически переведен (конечно, с потерей качества) на другой язык.

ЛИТЕРАТУРА

1. Feldman, Romen, and Ido Dagan. *Knowledge Discovery in Textual Databases*. Proc. Knowledge Discovery and Data Mining, pp. 112-117, Montreal, Canada, 1995.
2. Adolfo Guzman-Arenas. *Hallando los temas principales en un artículo en español*. Soluciones Avanzadas. Vol. 5, No. 45, p. 58, No. 49, p. 66, 1997.
3. Adolfo Guzman-Arenas. *Finding the main themes in a Spanish document*. Journal Expert Systems with Applications, Vol. 14, No. 1/2. Jan/Feb 1998, pp. 139-148.
4. Большаков И.А., Д.Дж. Кассиди, А.Ф.Гельбух. *Русский Роже: параллельные русский и английский варианты иерархического тезауруса с семантическими связями на базе пополненного тезауруса Роже*. Труды Международного семинара по компьютерной лингвистике и ее приложениям Диалог'95, Казань, май 1995.
5. Alexander Gelbukh. *Using a semantic network for lexical and syntactic disambiguation*. Proc. of Simposium Internacional de Computaciyn: Nuevas Aplicaciones e Innovaciones Tecnolygicas en Computaciyn, November 1997, Mexico.
6. Alexander Gelbukh, Igor Bolshakov, Sofia Galicia Haro. *Automatic Learning of a Syntactical Government Patterns Dictionary from Web-Retrieved Texts*. International Conference on Automatic Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA, USA, June 1998.
7. Alexandrov, M., P. Makagonov and K.Sboychakov. *Searching similar texts: some approaches to solution*. Proc. of International Academy of Informatics of UNO, Moldova, 1999.
8. Elworthy, D. *Does Baum-Welsh re-estimation help taggers?* Proc. of Fourth Conference on Applied Natural Language Processing, Stuttgart, Germany, 1994.