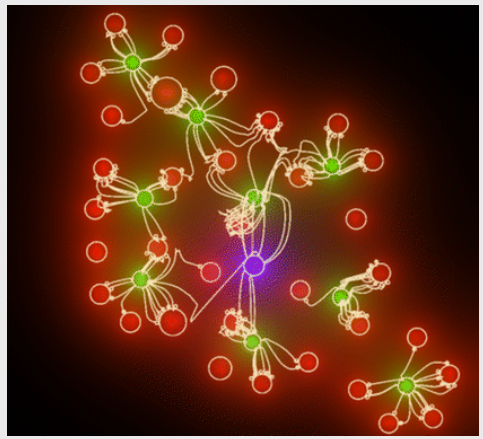
БИНЕЙРО



Система классификации текстов NNCS

СИСТЕМА "NNCS"

- Основа NNCS - модель представления текстовых документов в виде семантических векторов, получаемых с помощью специальной рекуррентной нейронной сети.

- "NNCS (Neural Network Classification & Search)

BiNeuro

Возможности системы

- Классификация текстов
- Информационный поиск
- Кластеризация текстов
- Ранжирование документов по семантической близости к запросу

Участие в РОМИП

- **В 2005 году компания «Бинейро» участвовала в РОМИП впервые**

- **Дорожка:**
  классификация веб-сайтов

Цели участия:

- ***апробация работы нового алгоритма семантического кодирования***

- ***сравнение результатов работы системы с результатами работы других систем подобного класса.***

Этапы работы системы

- Предварительная обработка текстов
- Создание семантической сети
- Получение семантических код-векторов текстов
- Классификация

Этапы работы системы

- Предварительная обработка текстов
- Создание семантической сети
- Получение семантических код-векторов текстов
- Классификация

Предварительная обработка

- Тексты WEB-сайтов, представленные в html-формате преобразуются в текстовые файлы с отбрасыванием всех тегов.
- Созданные таким образом текстовые файлы образуют корпус текстов, который используется в дальнейшем для формирования ассоциативной семантической сети.

Этапы работы системы

- Предварительная обработка текстов
- Создание семантической сети
- Получение семантических код-векторов текстов
- Классификация

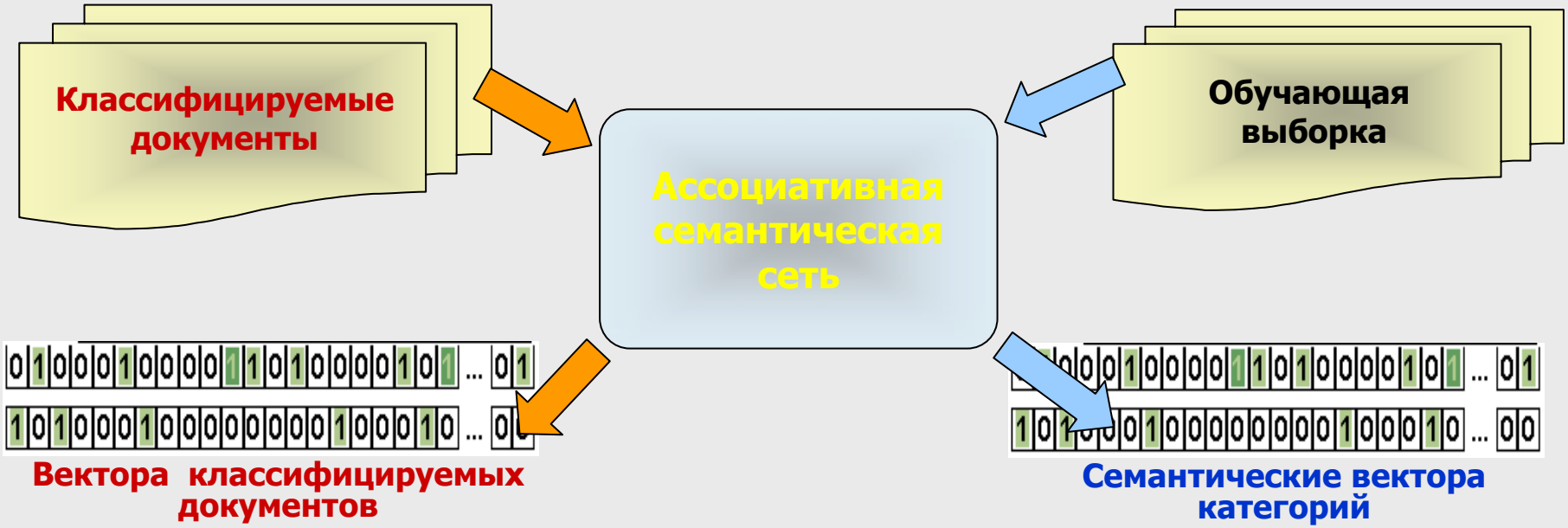
Создание семантической сети

- Вершины сети соответствуют концептам (часто встречающимся словам или устойчивым словосочетаниям, выражающим некоторое фундаментальное понятие).
- Словарь концептов генерируется с помощью готовых словарей и морфологического анализа корпуса текстов.
- Веса определяются специальным образом и зависят от статистики совместной встречаемости слов-концептов.

Этапы работы системы

- Предварительная обработка текстов
- Создание семантической сети
- Получение семантических код-векторов текстов
- Классификация

Получение семантических код-векторов документов



- Построенная ассоциативная семантическая сеть может быть ассоциирована с нейронной сетью Хопфилда с параллельной динамикой и с несимметрической матрицей обратных связей.
- На вход такой сети подаётся нормированный вектор, элементами которого являются статистические веса слов-концептов кодируемого документа.
- Окончательным код-вектором документа является положение равновесия сети. В силу определённого алгоритма формирования весов, оно всегда существует.

Этапы работы системы

- Предварительная обработка текстов
- Создание семантической сети
- Получение семантических код-векторов текстов
- Классификация

Классификация

Вектора классифицируемых документов

Семантические вектора категорий

ОБУЧЕНИЕ

Классификатор

документы, распределённые по рубрикам

- В результате работы нейронной сети каждому документу из обучающей и тестовой выборки ставится в соответствие действительный вектор, размерность которого совпадает с размерностью словаря концептов.
- Каждой категории ставится в соответствие семантический вектор документа, сформированного из всех документов данной категории, записанных подряд.

- Для каждого документа классифицируемой выборки по определённому правилу выбираются пять или меньше категорий, чьи семантические вектора наиболее близки к семантическому вектору документа в смысле стандартной евклидовой метрики.

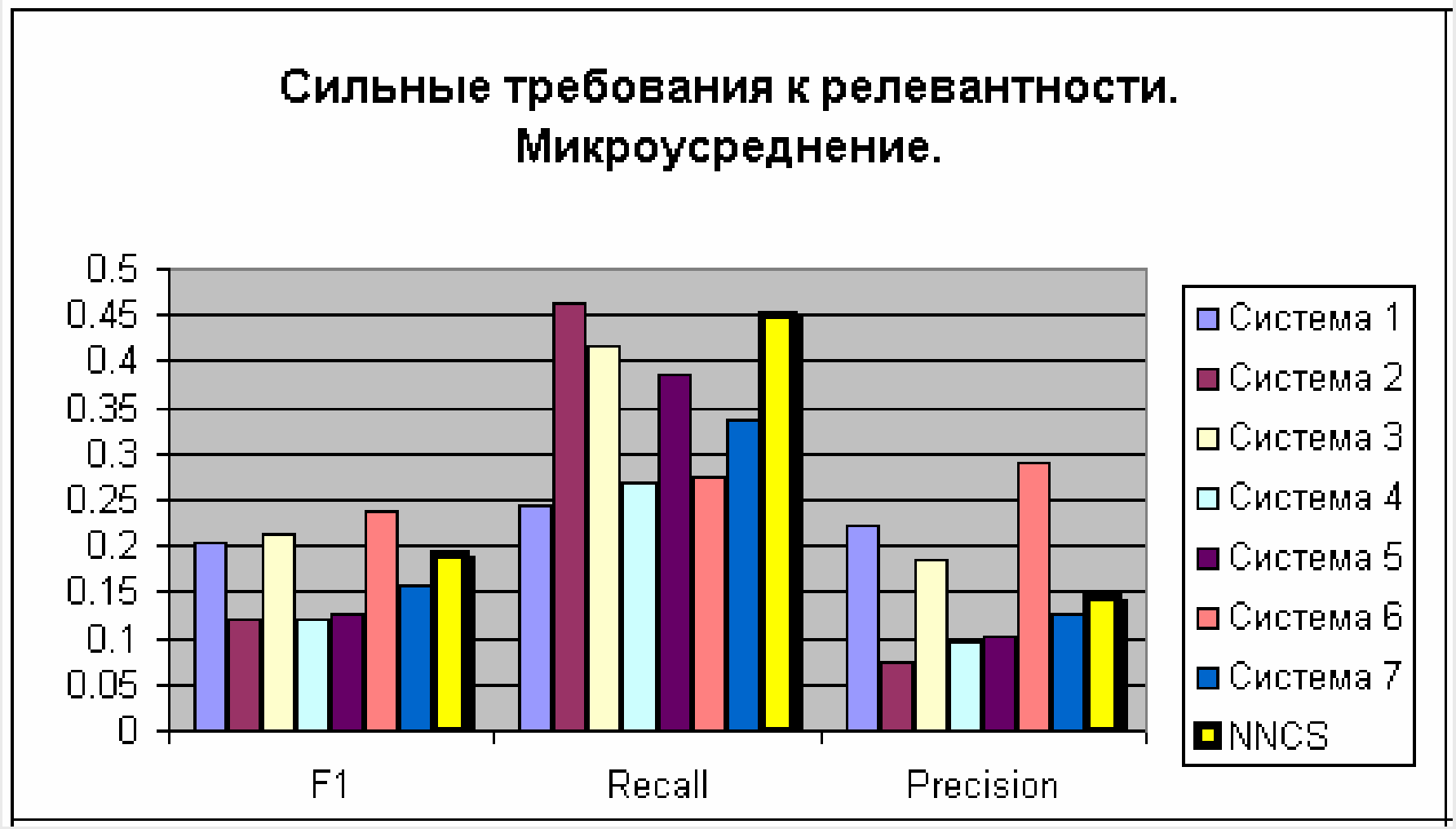
BiNeuro

Параметры теста для РОМИП

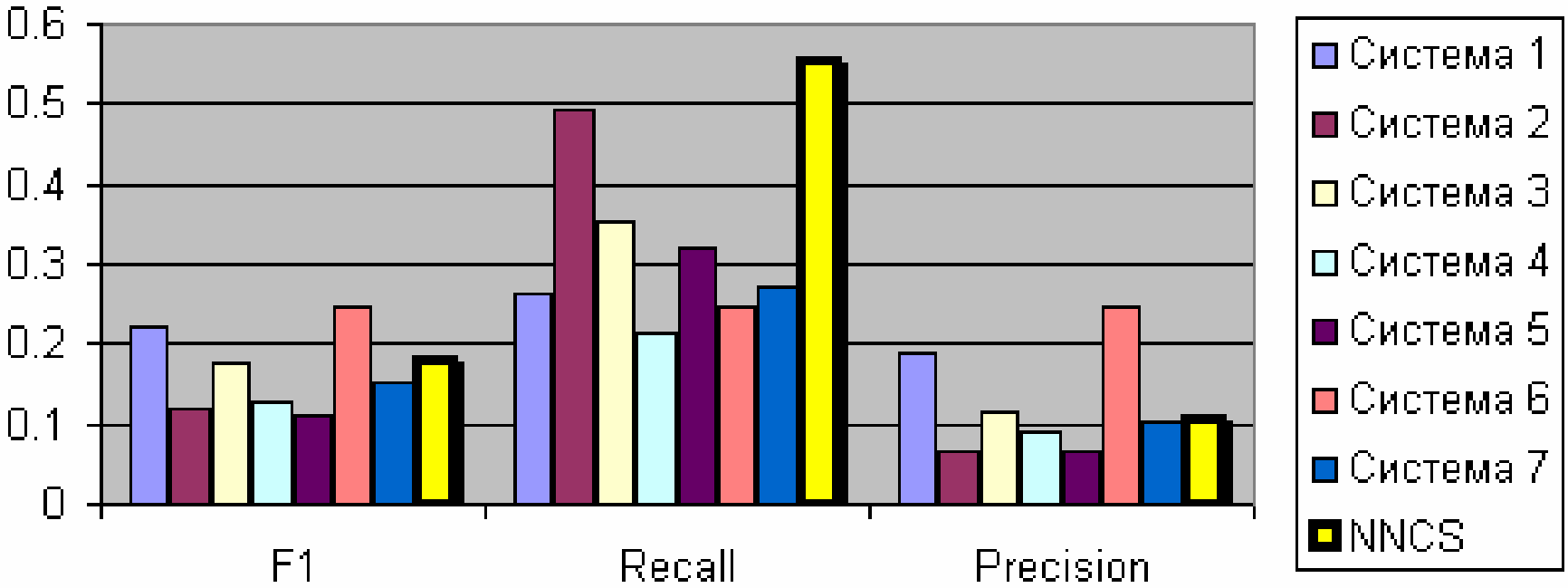
- Все программные компоненты системы NNCS, а также вспомогательные компоненты для работы с коллекциями DMOZ и narod.ru были реализованы на языке программирования «C#».
- В качестве корпуса для формирования нейронной сети были взяты все документы предоставляемой коллекции.
- Словарь концептов был сформирован из части встречающихся в корпусе прилагательных и существительных в нормальной форме. Его размер составил ~60000
- Для всех документов были получены семантические вектора, и для их классификации был использован описанный выше алгоритм.

Результаты

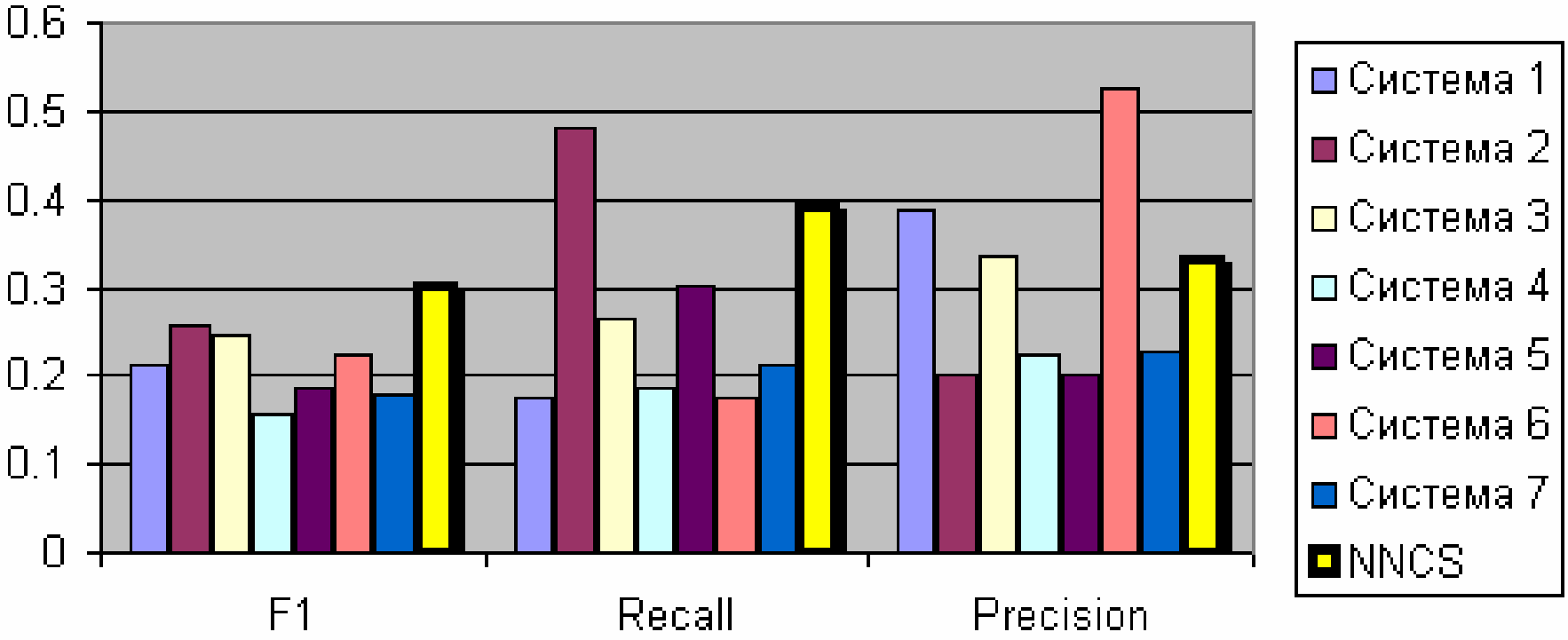
	AND	OR
F1 (macro average)	0.182	0.337
Recall	0.451	0.390
Precision (macro average)	0.108	0.270
Error	0.010	0.011
F1	0.191	0.302
Recall (macro average)	0.556	0.448
Accuracy	0.989	0.988
Precision	0.143	0.330



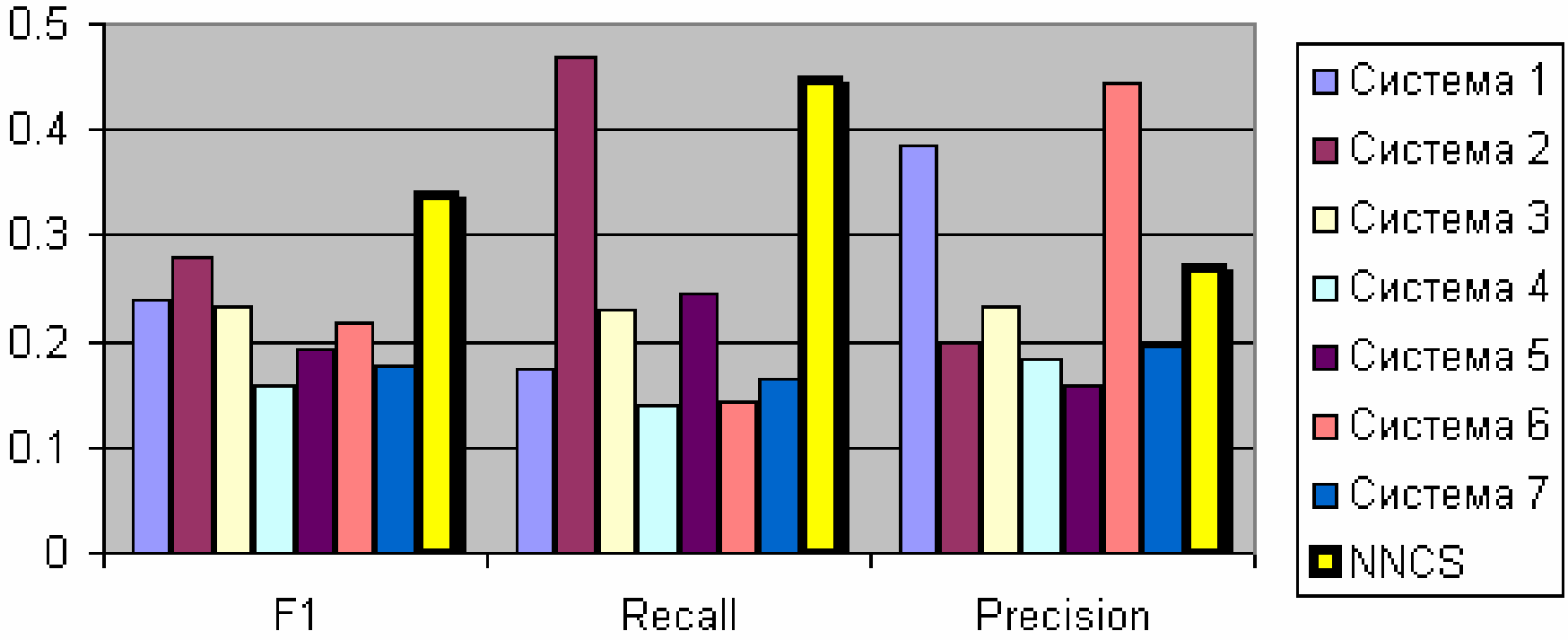
Сильные требования к релевантности. Макроусреднение.



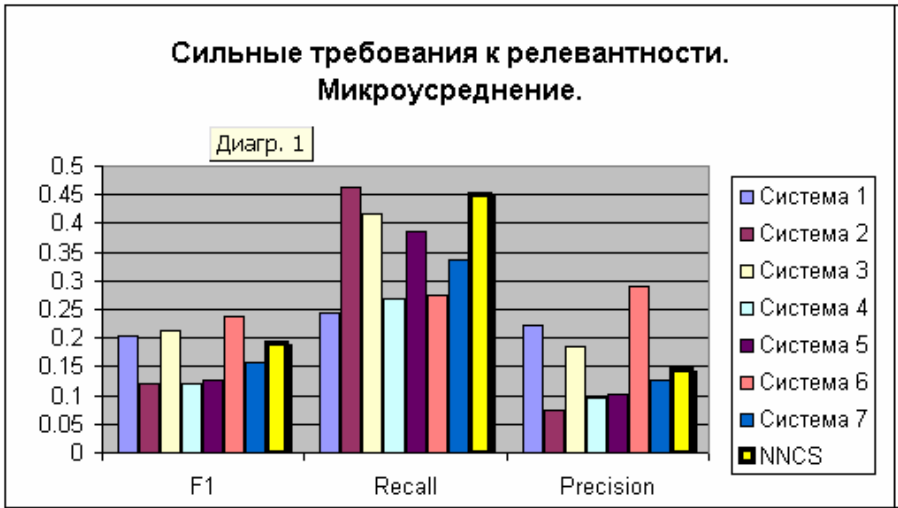
Слабые требования к релевантности. Микроусреднение.

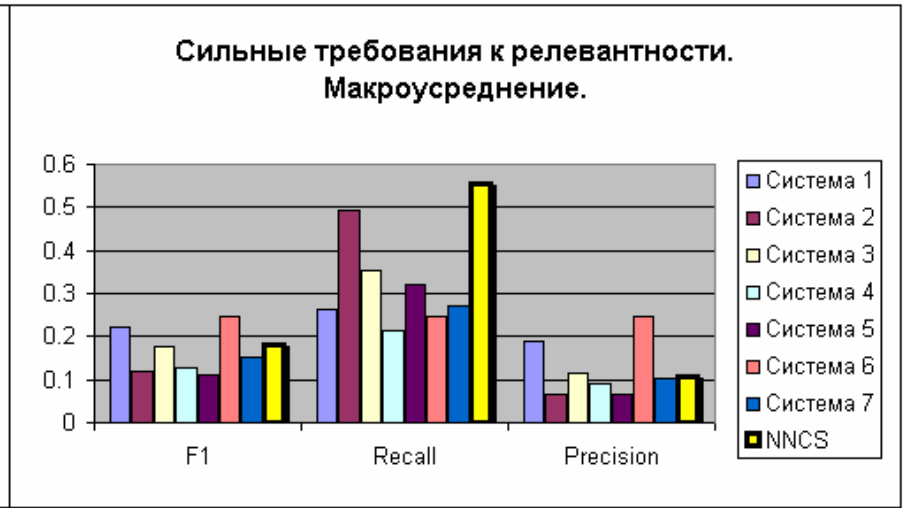


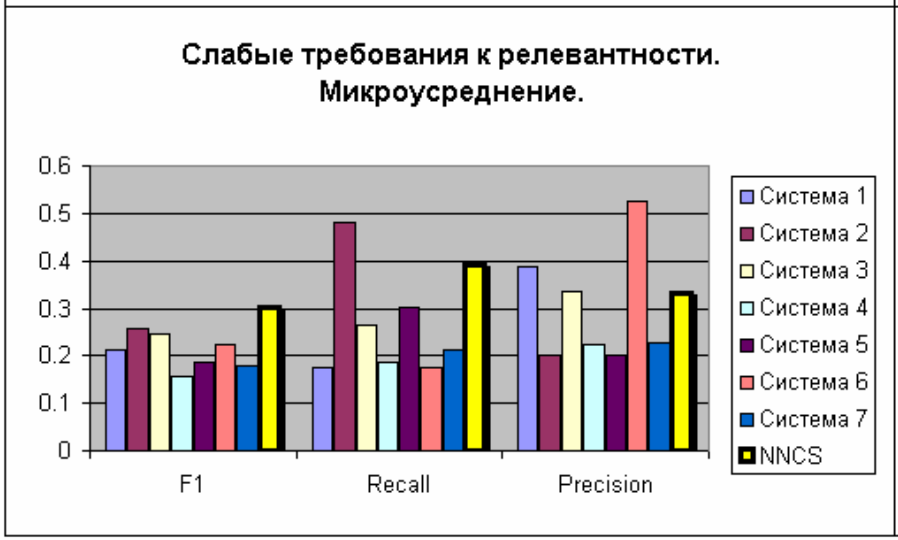
Слабые требования к релевантности. Макроусреднение.

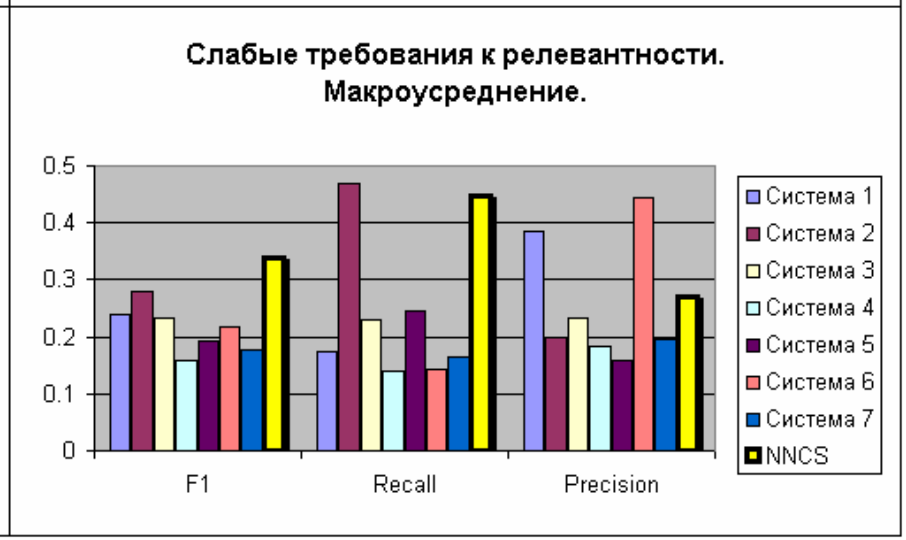


Сравнение результатов









ПАТЕНТ

На модель компанией «БИНЕЙРО» получен патент «Устройство кодирования семантики текстовых документов».


РОССЕЙСКАЯ ФЕДЕРАЦИЯ

ПАТЕНТ

НА ПОЛЕЗНУЮ МОДЕЛЬ

№ 45579

УСТРОЙСТВО ДЛЯ КОДИРОВАНИЯ СЕМАНТИКИ ТЕКСТОВЫХ ДОКУМЕНТОВ

Патентообладатель(и): *Открытое акционерное общество "Бинейро" (RU)*

Автор(ы): *Шмелев Александр Станиславович (RU)*

Заявка № 2005101306

Приоритет полезной модели: 09 февраля 2005 г.

Зарегистрировано в Государственном реестре полезных моделей Российской Федерации: 10 июня 2005 г.

Срок действия патента истекает: 09 февраля 2010 г.

Руководитель Федеральной службы по интеллектуальной собственности, патентам и товарным знакам

Спасибо за внимание!

mail@bineuro.ru