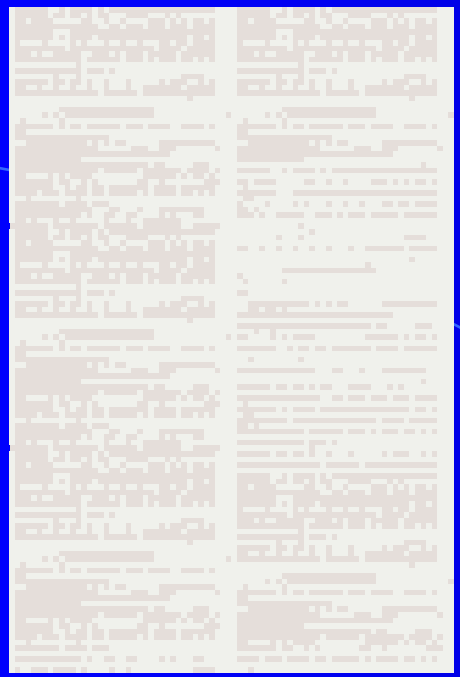


Галактика-Зум: Участие в семинаре РОМИП-2004

Антонов А.В., Козачук М.В., Мешков В.С.

Корпорация «Галактика»

Поисково-аналитическая система Галактика-Зум

- Предназначена для аналитической обработки больших объемов неструктурированной текстовой информации
- Сочетает возможности классической поисковой системы, системы сбора текстовых данных и системы аналитической обработки информации

Информационный Портрет – основа технологии Галактика-Зум

- Инфопортрет – список языковых инвариантов (слов и словосочетаний), отличающих выборку от прочих.

Участие в РОМИП - 2004

- Классификация веб-коллекции документов
- Тематическая классификация документов (нормативная коллекция)

Классификация веб-коллекции

Характеристики исходного массива данных

Общее количество сайтов	22 359
Общее количество веб-страниц	588 077
Общее количество рубрик	247

Характеристики базы данных Галактика-Зум (ГЗ)

Размер базы ГЗ (Мб)	8 288
Объем словаря слов	8 859 735
Число словомест	556 880 329
Объем словаря словосочетаний	3 120 022
Число мест словосочетаний	40 163 957

Классификация веб-коллекции

Фрагмент Информационного портрета классификационной рубрики «Компьютеры\Программирование»

Значение	*Вес*
ФУНКЦИЯ	55
ФАЙЛ	27.1
БАРХАТНЫЙ ПУТЬ	17.4
ЗНАЧЕНИЕ	16.8
СТРОКА	15.7
ОКНО	14.6
ПРИЛОЖЕНИЕ	14
ХОСТИНГ	13
УКАЗАТЕЛЬ	12.4

Значение	*Вес*
ПАРАМЕТР	12.1
ВОЗВРАЩАТЬ	11.1
ОПЕРАЦИОННАЯ СИСТЕМА	10
ИДЕНТИФИКАТОР	9.95
ФАЙЛОВАЯ СИСТЕМА	9.45
ПЕРЕМЕННЫЙ	7.98
КЛАСС	7.97
СИМВОЛ	7.09

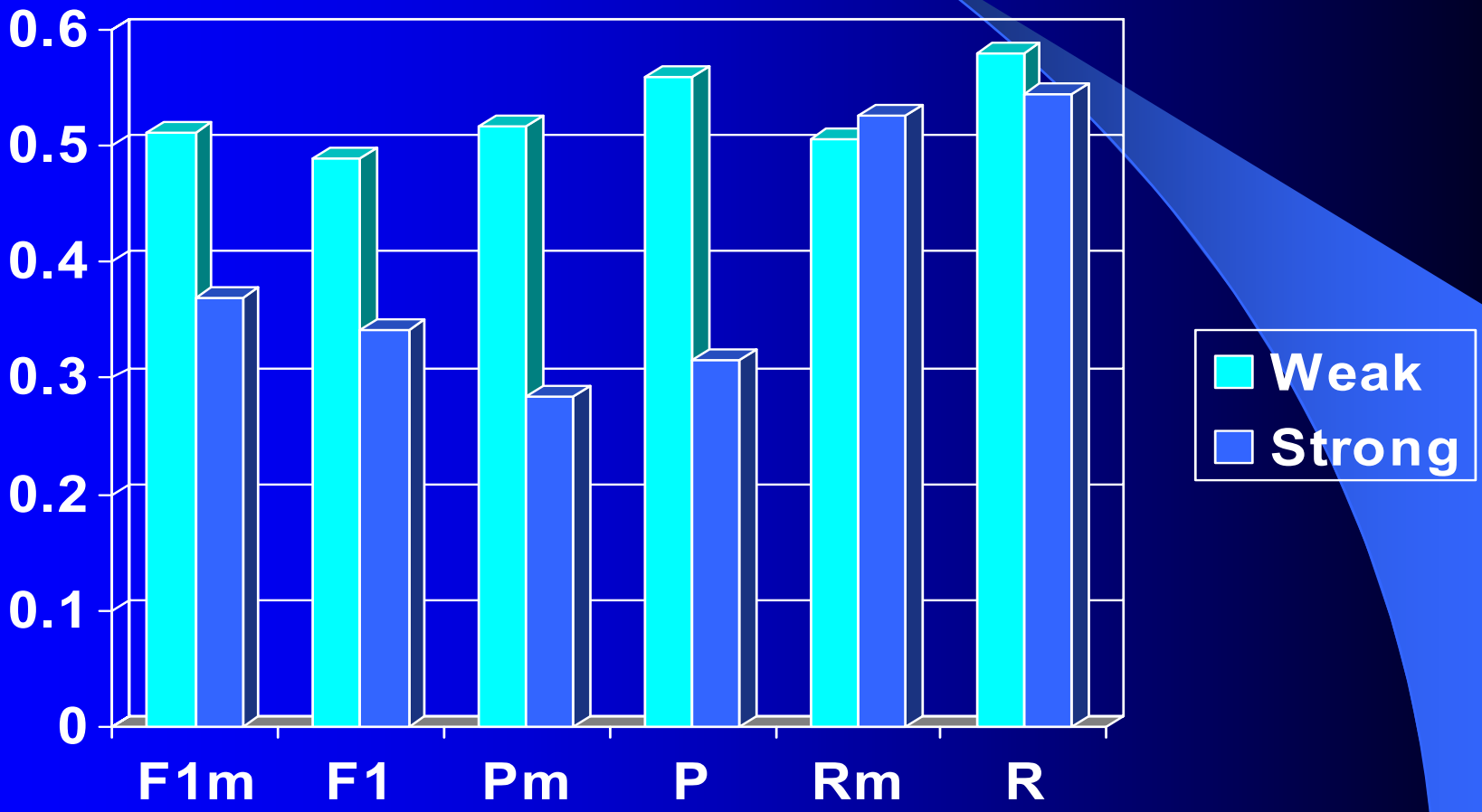
Классификация веб-коллекции

Оценки качества классификации

Оценка	***weak***	***strong***
F1 (macro average)	0.511	0.369
Recall	0.579	0.544
Precision (macro average)	0.517	0.285
Error	0.009	0.008
F1	0.49	0.341
Recall (macro average)	0.505	0.526
Accuracy	0.991	0.991
Precision	0.559	0.316

Классификация веб-коллекции

Оценки качества классификации



Тематическая классификация

Характеристики исходного массива данных

Общее количество документов	60 294
Общее количество рубрик	183
Количество классифицированных рубрик (содержат хотя бы один документ)	162

Характеристики базы данных Галактика-Зум (ГЗ)

Размер базы ГЗ (Мб)	816
Объем словаря слов	1 017 360
Число словомест	87 073 269
Объем словаря словосочетаний	340 879
Число мест словосочетаний	10 171 889

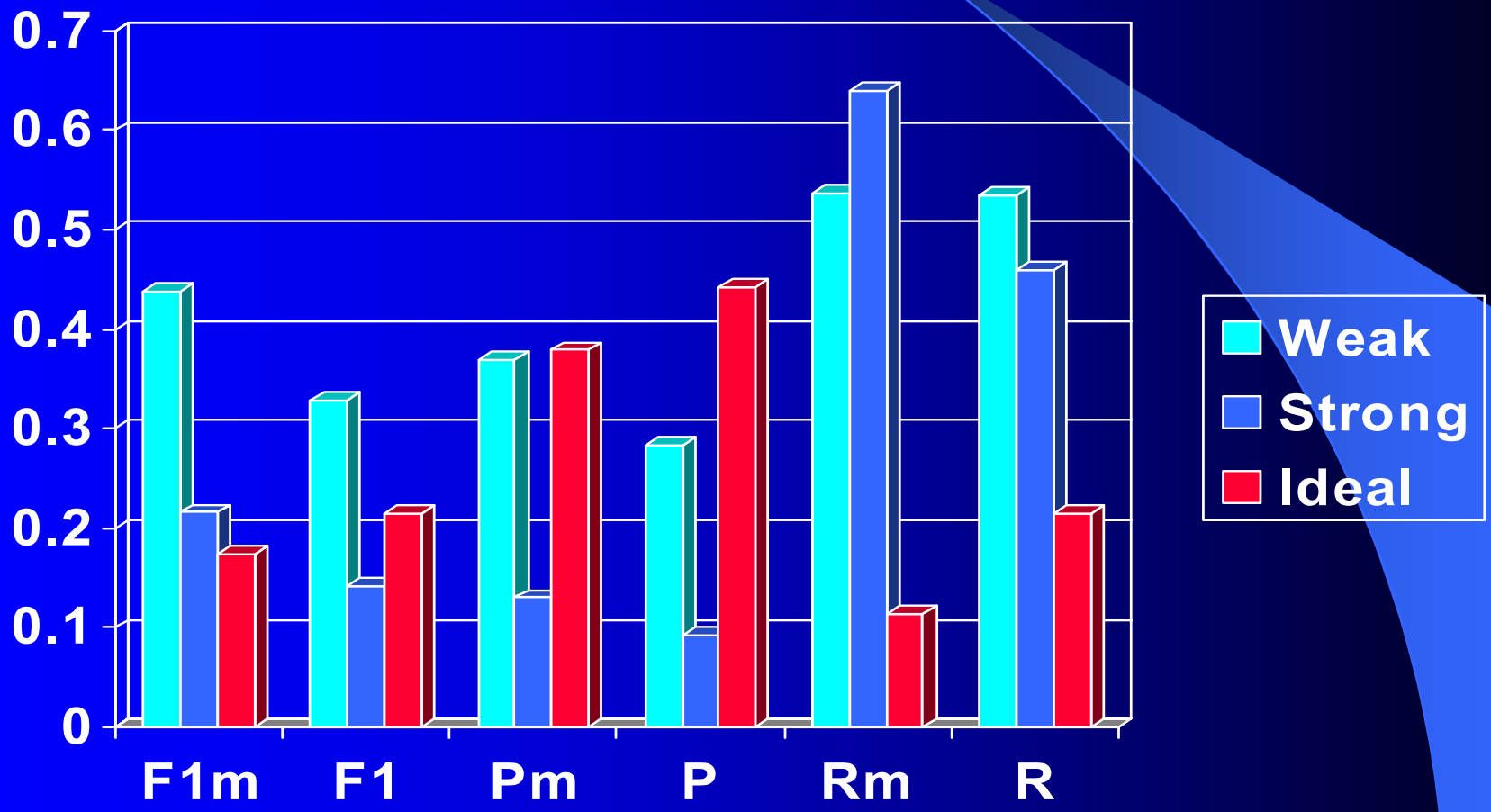
Тематическая классификация

Оценки качества классификации

Оценка	*weak*	*strong*	*Ideal*
F1 (macro average)	0.438	0.216	0.174
Recall	0.535	0.46	0.214
Precision (macro average)	0.37	0.13	0.379
Error	0.03	0.028	0.033
F1	0.328	0.141	0.214
Recall (macro average)	0.537	0.64	0.113
Accuracy	0.97	0.97	0.967
Precision	0.283	0.092	0.442

Тематическая классификация

Оценки качества классификации



Опыт участия в РОМИП-2004

- Независимая оценка результатов исследований
- Возможность познакомиться с результатами других исследований в данной области

Благодарю за внимание

Демо-сайт
http://zoom.galaktika.ru