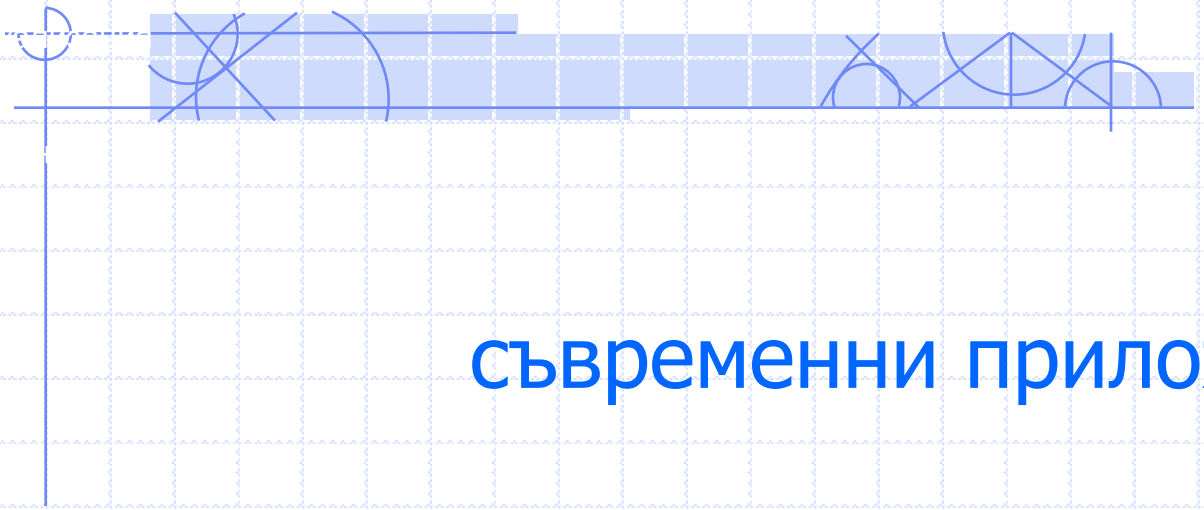
Информационни системи



съвременни приложения

Теми

- Въведение
- Съвременни приложения
- Визуализация на информация
- Тенденции

Информационни системи

Организация на хора и компютърни технологии за управление на информация

Програмни системи за събиране, съхраняване, обработване, извличане и разпространение на информация

Компоненти

- база от данни
- управляващи програми
- потребителски интерфейс
- комуникационна среда

Видове

- според областите на приложение: всички области
- според източника на данни: хора или машини
- според предназначението: DP, MIS, DSS, KBS, …

Съвременно развитие

- Технологична среда
  - Интернет – информационна вселена без граници
  - мултимедия
  - средства за визуализация
  - изчислителна мощност на компютрите
- Социална среда
  - разширяване на приложните области
  - повишена технологична култура на потребителите
  - повишени изисквания към функционалността на системите

Съвременни приложения

- Извличане на информация
  - Information Retrieval
- Анализ на данни
  - Data Analysis
- Откриване на информация
  - Data Mining
- Извличане на данни в мултимедия
  - Multimedia Mining
- Географски информационни системи
  - GIS
  - навигационни средства
- Други

IR

- Системи за събиране, съхраняване, организация и достъп до елементи информация
- Мотивация за развитие
  - Internet
    - универсално хранилище на човешкото знание и култура
    - безплатен източник на информация
- Проблеми
  - нерегламентирани заявки за извличане
  - разнородни информационни източници
  - огромно количество
  - продължително търсене

IR

- **Причини за проблемите**
  - липса на единен модел на данните, съдържащи се в Интернет
  - липса на качествени описания и структури
  - изоставане на софтуерните технологии в сравнение с потребителските фактори
- **Резултат**
  - поставяне на IR в центъра на изследванията
  - разграничаване на DR и IR: извличане на контекстно зависима информация с високи потребителски качества
  - фокус върху нуждите на потребителя
- **Съвременни изследвания**
  - моделиране на данни
  - методи за филтриране, класификация и категоризация на документи
  - потребителски интерфейси
  - езици за формиране на заявки
  - визуализация на информацията
  - системни архитектури
  - и др.

IR vs. DR

- **DR**
  - намиране на документи, съдържащи определени ключови думи
  - добре дефинирана семантика
  - нетърпимост към грешки
- **IR**
  - информация относно тема
  - недефинирана семантика
  - толеранс на грешките
- **Информационна система IR**
  - интерпретира съдържанието на намерените документи
  - ранжира намереното по отношение на потребителски критерии
  - оценява приложимостта на намерените документи

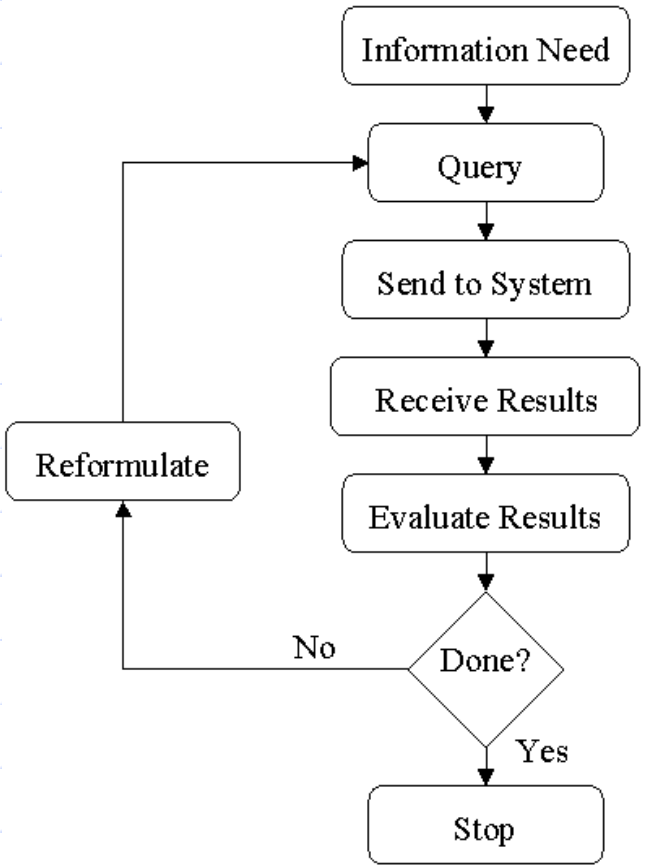
IR

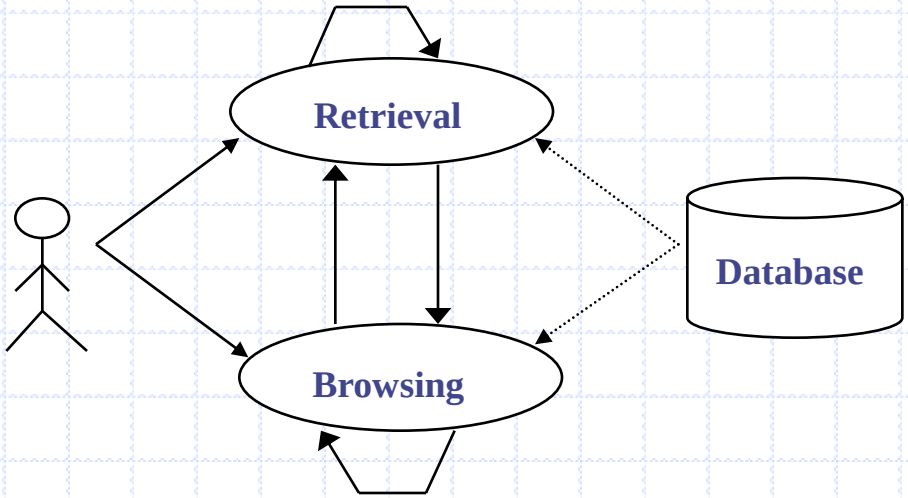
- Основни информационни единици
  - текстови документи
- Съдържание на документите
  - синтаксис
  - семантика
  - структура
- Метаданни за документи
  - дескриптивни
  - семантични

Модели на търсене в IR

- Прост интерактивен модел
  - дълги списъци от намерени документи без оценка на приложимостта
  - статична цел на потребителя
  - итеративно-настройващи се заявки
  - използва се от web search machines
- Самообучение на потребителите
  - използване на хипервръзки и навигация в търсенето (near-miss)
  - `berry-picking' model
    - актуализация на целите на потребителя
    - обобщаване на резултати от множество под-заявки
    - главен резултат: акумулираното знание, получено по време на търсенето, не в крайната извадка

Модели на търсене в IR





Видове търсене в IR

- **разглеждане**
  - ненасочено изследване на информационни структури
  - следва се от селекция
- **запитване**
  - получаване на ново (под) множество от информационни елементи, не събирани заедно досега
  - следва се от разглеждане
- **селектиране**
  - избор от организирана информация
  - използва се като резултат или за формулиране на запитване
- **сканиране**
  - целенасочено изследване на заглавия, термини, категории и др.
- **навигация**
  - сканиране + селектиране

Формулиране на заявки

- Методи
  - Избор на ресурсна колекция, метаданни или информационно множество, отговарящо на критериите за приложимост
  - Специфициране на думи, фрази, дескриптори за сравнение
- Средства
  - команден език
  - попълване на формуляри
  - избор от меню
  - директна манипулация
  - естествен език

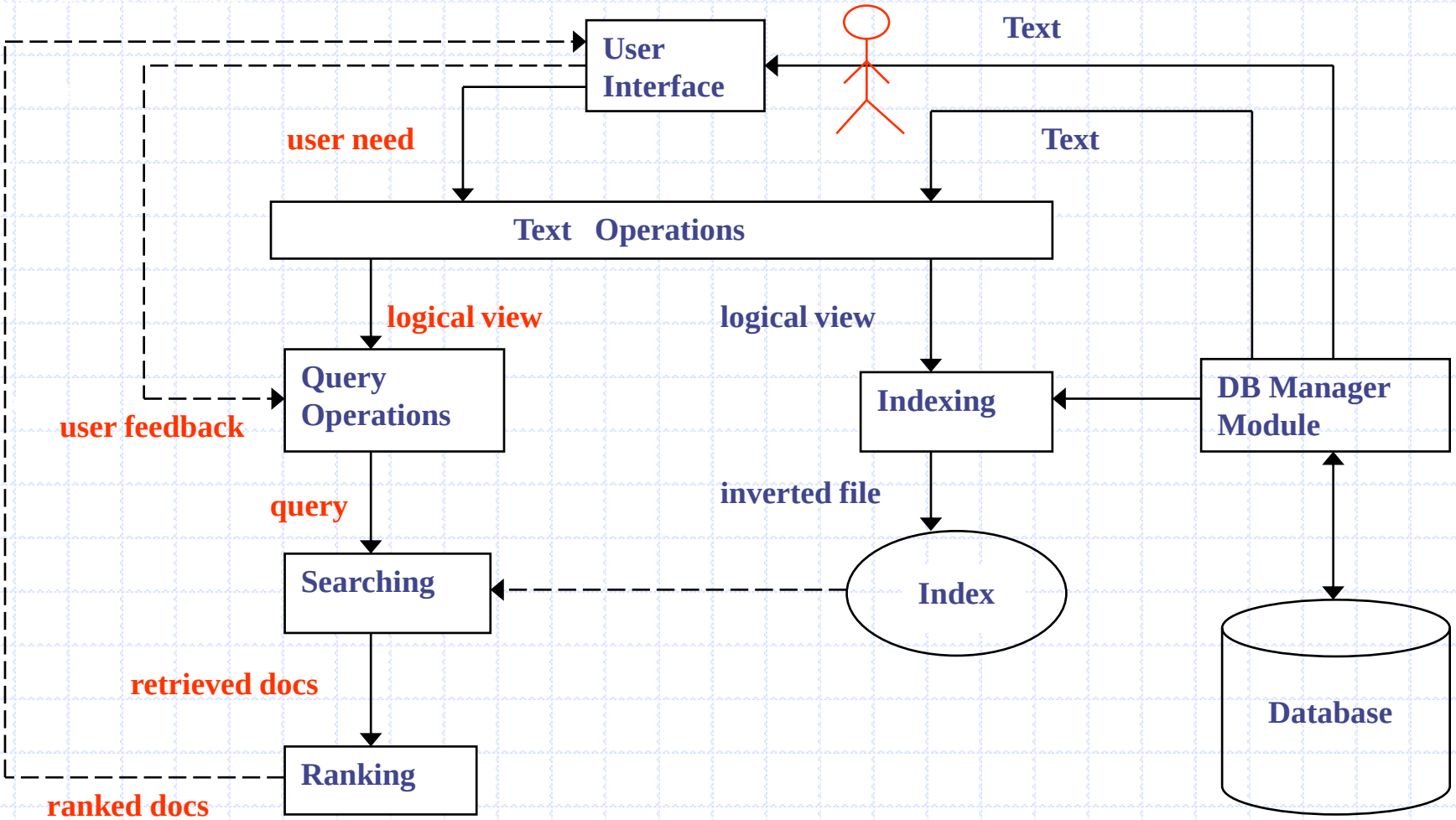
Езици за заявки

- Дума и фраза
  - дизюнкция или конюнкция ?
  - фасети и етикети за улеснение на потребителя
  - приложимостта на резултата зависи от алгоритъма на ранжиране
    - статистически
      - тегло
      - вероятност
      - процент
- Естествен език
  - потребителят контролира приложимостта
  - необходимо е да познава разпознаваемите команди

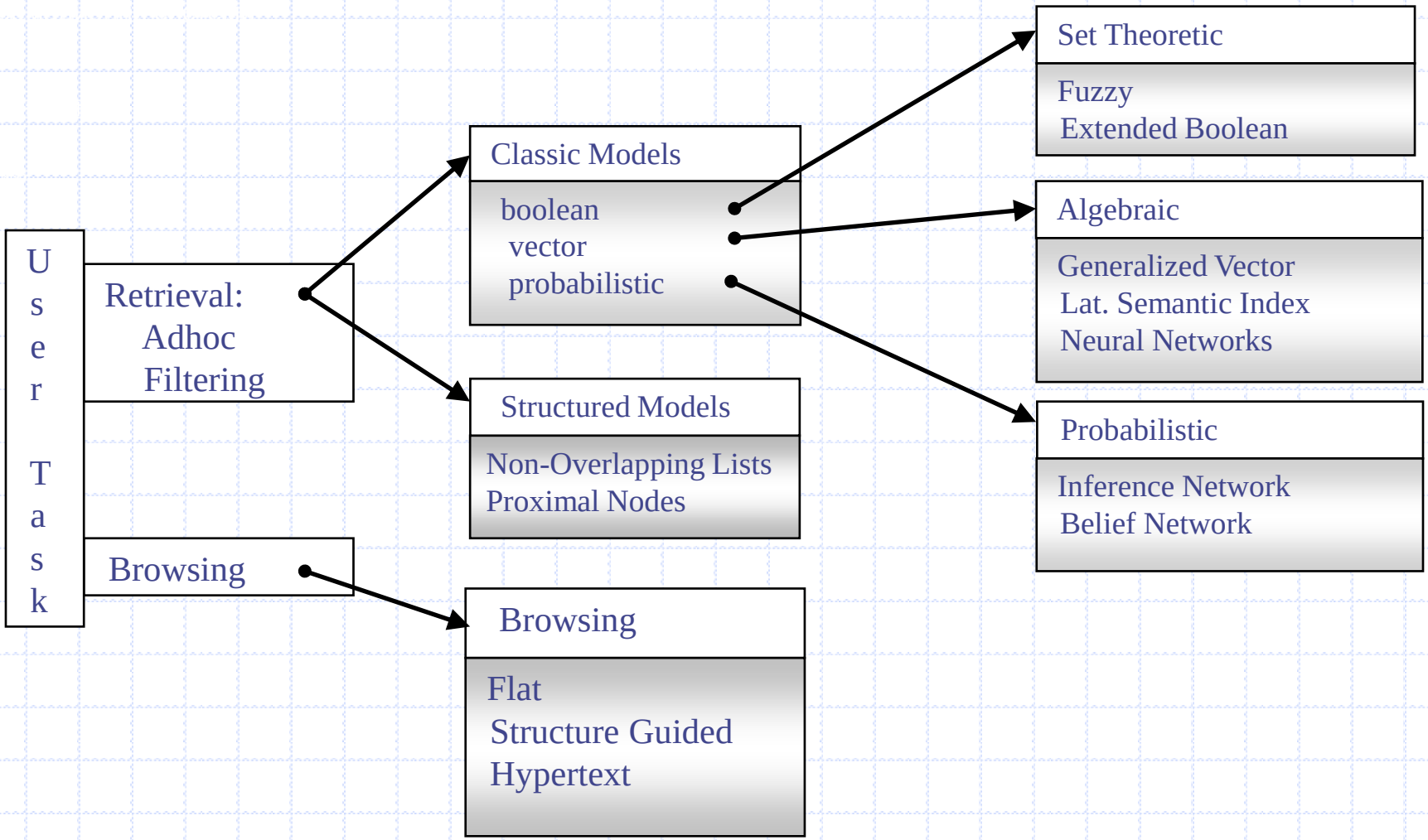
Естествени езици

- Въпроси и отговори
  - от синтаксиса на въпроса се извлича информация за свойствата на отговора (напр. каква част от изречението е)
- FAQ finder
  - установяване на съответствие между двойки въпрос и отговор
- Предефинирани типове въпроси
  - идентифициране на типа въпрос на потребителя и последващо перифразиране (доуточняване)
- Свободен текст на въпросите
  - алгоритъм за разделяне на използваните думи и оценяване на тяхното значение

Процес на извличане на информация

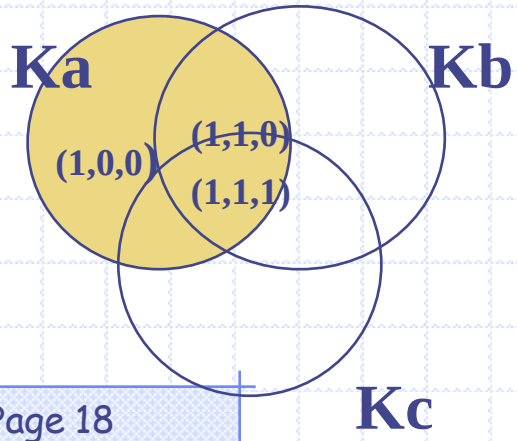


Видове модели на IR



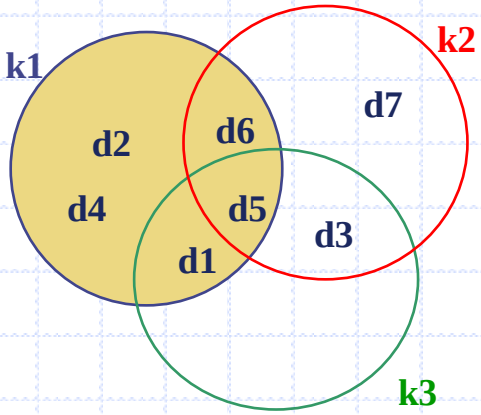
Примери

- Индекси на документите
- Заявки за намиране на съвпадение на индекси
- Булево търсене в множество
  - функция на принадлежност {0,1}
- Размито множество
  - функция на принадлежност [0,1]
  - степен на приложимост на документа



Примери

- Векторен модел
  - степен на приложимост на документа



	k1	k2	k3	q • dj
d1	2	0	1	5
d2	1	0	0	1
d3	0	1	3	11
d4	2	0	0	2
d5	1	2	4	17
d6	1	2	0	5
d7	0	5	0	10
q	1	2	3	

Специализирани приложения

- Digital Libraries
  - Многоезични документи
    - речници
  - Мултимедийни документи
    - синхронизация на потоци информация
    - намиране по метаданни
    - визуални езици за заявки
    - *Query By Image Content*
  - Структурирани документи
    - приложени една или повече структури на данните
    - описания чрез SGML
  - Разпределени документи
    - във физическо или логическо пространство
- Намиране на родственици!

Разпределени документи

- Federated search
  - изпращане на заявки към различни сървъри и обобщаване на резултатите в единен формат

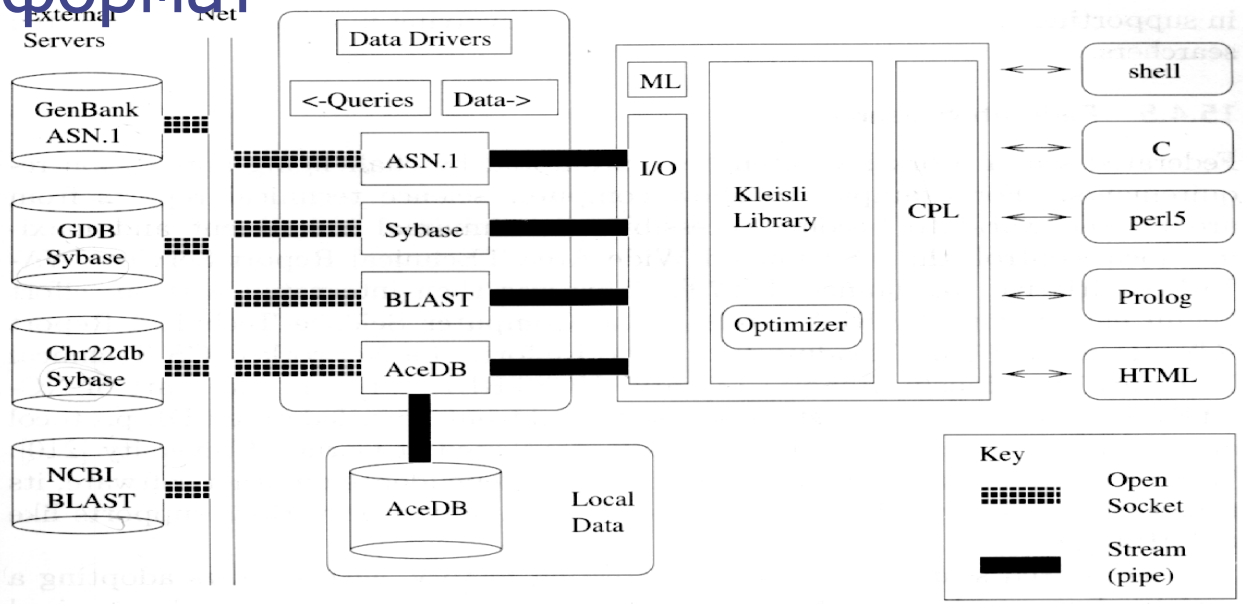


Figure 15.2 Architecture of the BioKleisli system (adapted from [829, 128]).

Стандарти за IR

- Комуникационни протоколи
  - Z39.50
  - WAIS
  - Dienst
- Стандарти за данни
  - Resource Description Framework (RDF)
    - разделяне на обект и описание
  - Text Encoding Initiative (TEI)
    - комбиниране на данни и метаданни
- Стандарти за метаданни
  - MARC
  - Dublin Core - 15 основни елемента за описание на цифров обект по отношение на:
    - content (Title, Subject, Description, Source, Language, Relation, Coverage)
    - intellectual property issues (Creator, Publisher, Contributor, Rights)
    - digital objects (Data Type, Format, Identifier)
  - Warwick Framework
    - пакети и връзки между тях

OLAP

- On-Line Analytical Processing
- Извличане на информация, подпомагаща вземането на решения
- Позволява проверка на хипотези
  - IR показва какво се съдържа в базата данни
  - OLAP показва защо съществуват определени зависимости
- Прилага се на предварителните етапи на разкриване на информация и знания

Data Mining

- Определение
  - процес по анализиране на големи масиви от данни (тера-байтове)
  - с цел разкриване на зависимости, които могат да помогнат
  - при построяване на прогнозни модели
  - за управление на вземане на решения

Определения

- не-тривиално извличане на предварително неизвестна и потенциално полезна информация от данни
- наука за откриване на полезна информация от големи множества или бази от данни
- подобно на “Изкуствен интелект” - обобщаващ термин за дейности, свързани с откриване на информация в широк смисъл и контекст
- относително нова технология за анализ на данни, която се основава на традиционни изчислителни техники, като
  - статистика
  - машинно обучение
  - разпознаване на образи
  - и др.

Приложения

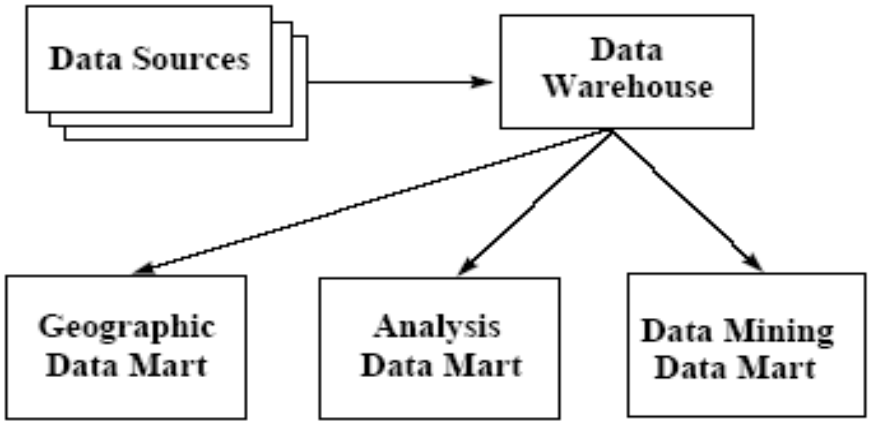
- Приложни области
  - В бизнеса
    - customer life cycle
    - електронна търговия
    - маркетинг
    - прогнозиране
  - В ГИС и демография
  - В научни изследвания
  - и др.
- Условия за успешно приложение
  - коректно дефиниране на проблем
  - използване на подходящи данни

Процес

- Събиране, анализ и подбор на данни
- Предварително описание и оценка със статистически параметри
- Визуализация и избор на зависимости за изследване
- Построяване на прогнозен модел
- Тестване на модела
- Верификация на модела

Data Warehouse и Data Mart

- Data Warehouse
  - контейнер за данни от различни източници и в различни формати
- Data Mart
  - извадка от данните, подходяща за конкретно изследване; read-only database

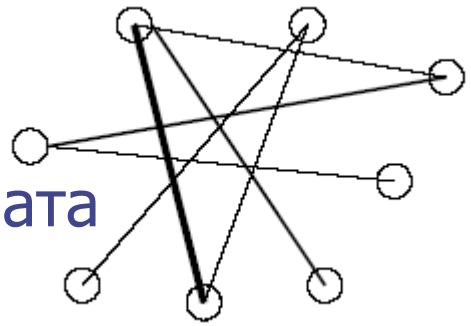


Предварително описание

- Клъстеризация
  - разделяне на данните в относително самостоятелни групи
  - групите не са известни предварително
  - използват се за класифициране на нови данни

Предварително описание

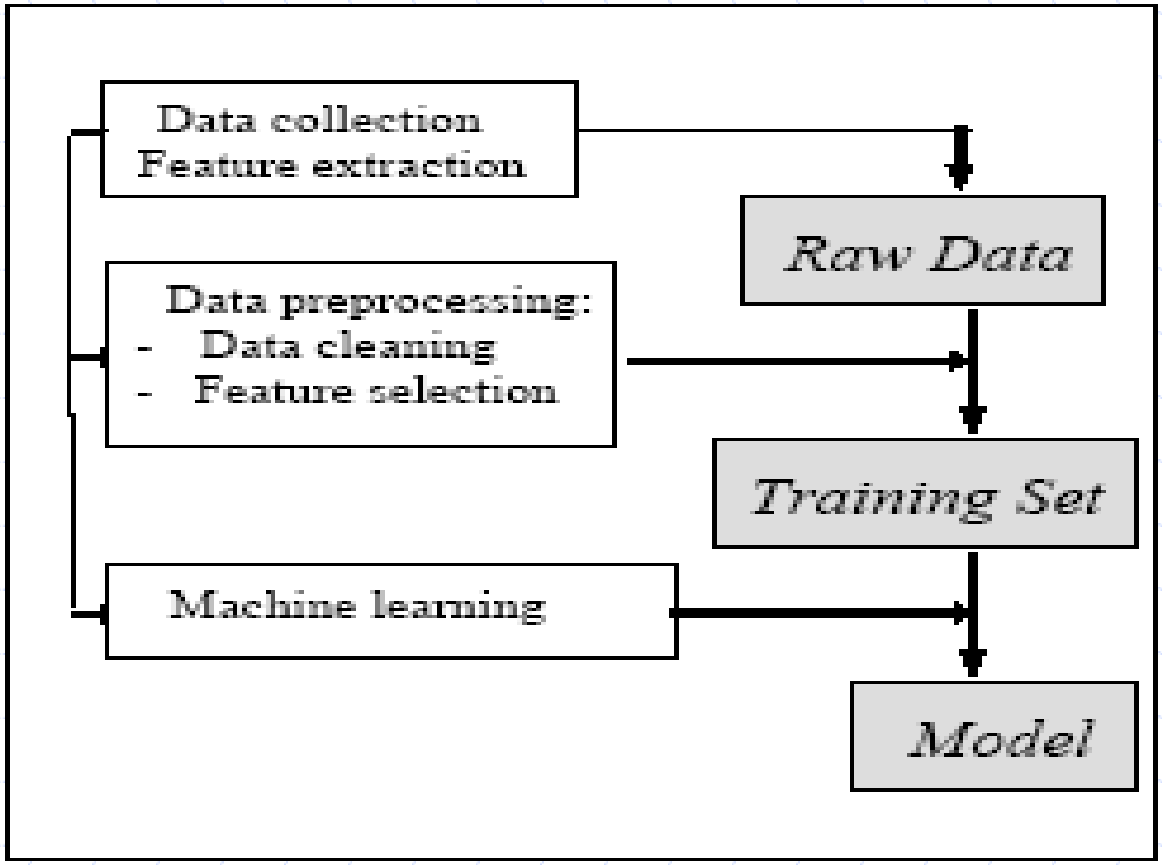
- Анализ на връзките
  - идентификация на отношения между данните
  - два подхода
    - откриване на асоциации
    - откриване на последователности
  - откриване на правила
    - дефиниране на отношения
    - оценка на валидността на правилата
      - confidence, lift
  - графично представяне на връзките



Построяване на модел

- Цел на моделирането
  - описание на шаблони и отношения между данните, които могат да се използват за прогнозиране
- Етапи
  - дефиниране на проблем
  - избор на тип прогнозиране
    - класификация
    - регресия (стойност) или f(t)
  - избор на тип модел, напр.
    - дърво на решенията
    - невронна мрежа
  - построяване / избор на алгоритъм, напр.
    - back-propagation
    - CART or CHAID
  - програмиране
    - избор на софтуер

Построяване на модел



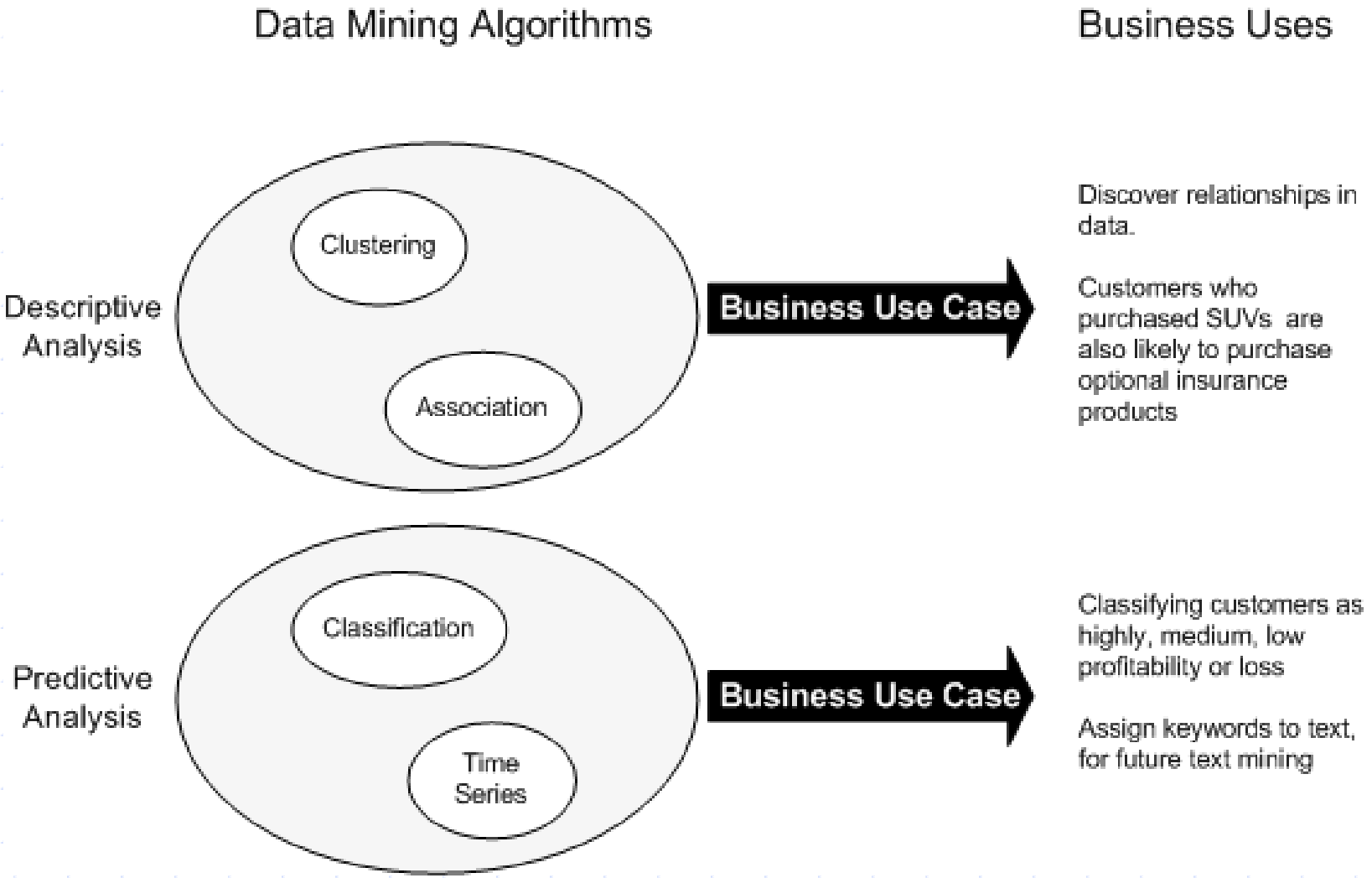
Видове алгоритми

- **Алгоритми за класификация и прогнозиране** на една или повече дискретни променливи (атрибути) в зависимост от други, входни атрибути
- **Регресионни алгоритми** за прогнозиране на непрекъснати величини, в зависимост от други
- **Алгоритми за сегментиране** – разделяне на данните в групи с подобни свойства - клъстери
- **Асоциативни алгоритми** – за откриване на корелации между входните данни - откриване на правила за асоциации
- **Анализ на повтарящи се последователности** от данни

Видове алгоритми

- Невронни мрежи
  - за голям брой фактори и отношения
- Дърво на решенията
  - правила за причисляване към клас или стойност
- MARS
  - Multivariate Adaptive Regression Splines
- K-nearest neighbor
- Time Series Algorithm
- Logistic regression
- Discriminant analysis
- GAM - General Additive Models
- Genetic algorithms
- и др.

Примерна употреба



Примерна употреба

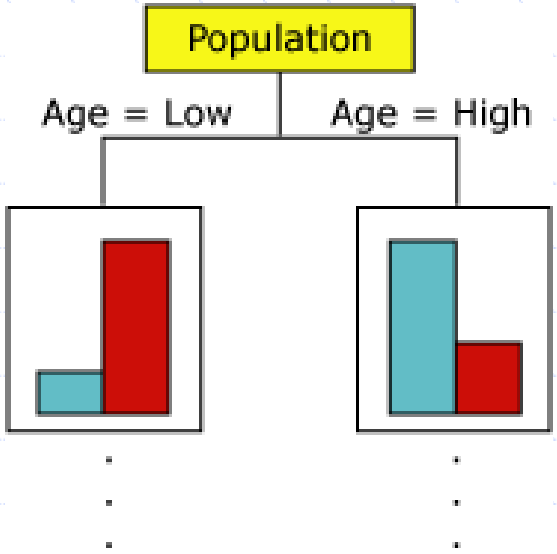
Задача	*Алгоритъм*
Прогнозиране на дискретни величини *напр. дали получателите на реклама ще купят стоката*	Дърво на решенията Алгоритъм на Бейс Невронни мрежи
Прогнозиране на непрекъснати величини *напр. предвиждане на продажбите за следващата година*	Дърво на решенията Time Series
Прогнозиране на последователности *напр. анализ на избора на път в сайт*	Sequence Clustering
Откриване на подобие в транзакции *напр. анализ на потребителска кошница за добавяне на стоки в нея*	Асоциации Дърво на решенията
Откриване на подобие в данни *напр. сегментиране и анализ на демографски данни за откриване на значимост на атрибути*	Кластеризация

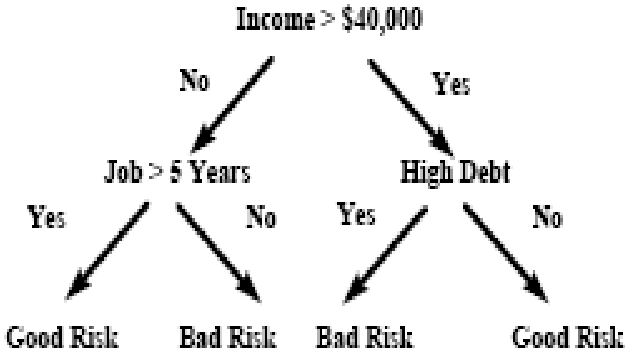
Дърво на решенията

- Алгоритъм за класификация и регресия, приложим за прогнозно моделиране на дискретни и непрекъснати атрибути
- Моделът съдържа ключов атрибут (колона), входни колони и колона, подлежаща на прогнозиране
- За всяка изходна колона, подлежаща на прогнозиране се строи дърво на решенията
- Дървото се разклонява, когато се открие значима корелация между входен атрибут и прогнозиран атрибут

Дърво на решенията

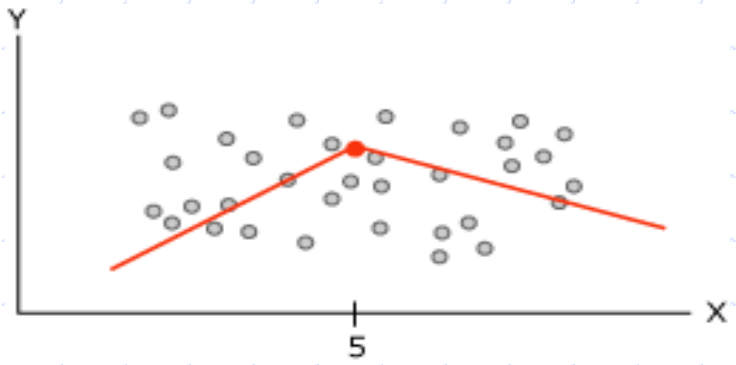
- За прогнозиране на дискретни атрибути:
  - прогнозите се основават на отношенията между входните колони данни
  - ако е установена тенденция или атрибут, открояващ тенденция, то те се използват за прогноза на стойността на друг, изходен атрибут
  - пример: атрибутът "възраст" може да се използва за прогнозиране на атрибут "потенциален купувач на велосипед"

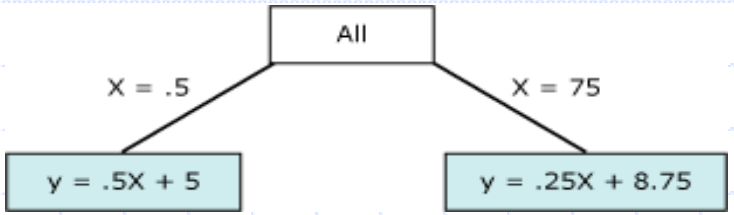




Дърво на решенията

- За прогнозиране на непрекъснати атрибути:
  - използва се линейна регресия за определяне на точката на разделяне на дървото
  - това е точката на пресичане на апроксимиращите прави линии





Алгоритъм на Бейс

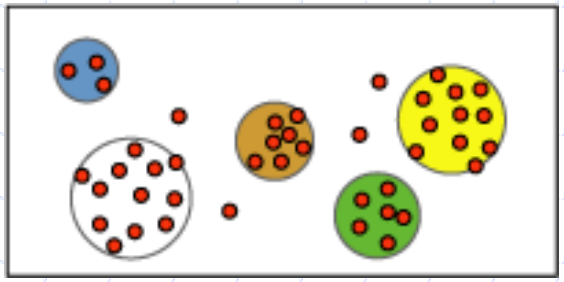
- Алгоритъм за класификация и прогнозиране
- Изчислява се условната вероятност между входните и изходните атрибути
- Атрибутите се приемат за независими помежду си
- Използва се за предварително изследване и оценка на данните
- За всяка възможна стойност (състояние) на изходен атрибут (резултат) се изчислява вероятността (разпределението) на всяка стойност (състояние) на входните атрибути

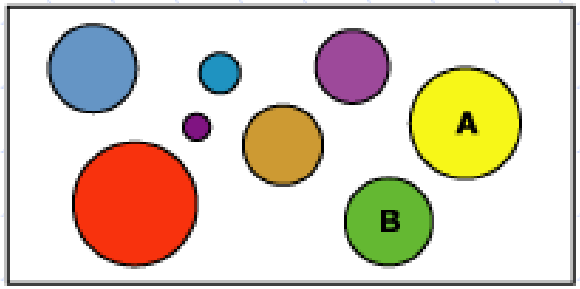
Пример

Attributes	States	Population ... Size: 18484	0 Size: 9352	1 Size: 9132	missing Size: 0
Age	38 - 43 29 - 34 43 - 48 Other ...				
Commute Distance	0-1 Miles 2-5 Miles 1-2 Miles Other ...				
Education	Bachelors Partial College High School Other ...				
Marital Status	M S Missing				
Number Cars Owned	2 1 0 Other ...				
Number Children At Home	0 1 2 Other ...				
Occupation	Professional Skilled Manual Management				

Клъстеризация

- Отделяне в групи по подобие
- За изследване на прилики и аномалии
- Изследва зависимости между входните данни и откритите клъстери
- Итерационен процес за оптимизиране на клъстерите
- Не е необходим изходен атрибут
- Методи за кластеризация
  - вероятностни – проверява се вероятността дадена стойност да принадлежи на клъстер - *Expectation Maximization (EM)*
  - оценява се разстоянието на стойност до клъстер и се определя най-малкото разстояние - *K-Means*

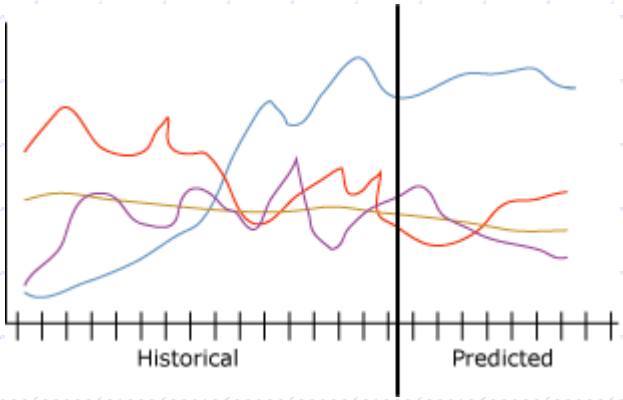




A = Commuters who drive to work
B = Commuters who bicycle to work

Времеви редове

- Time Series
- За прогнозиране на непрекъснати величини
- Основава се на тренд, извлечен по време на анализ на входните атрибути
- Елементи на модела
  - времев ред на данните
  - времев ред на прогнозата
  - комбинация от двата
  - *case series* – атрибутът, който разграничава точките в серията, напр. дата, йерархии от периоди, празници и други специални моменти от времето
- Използват се предварително известни резултати за ръководене на прогнозирането

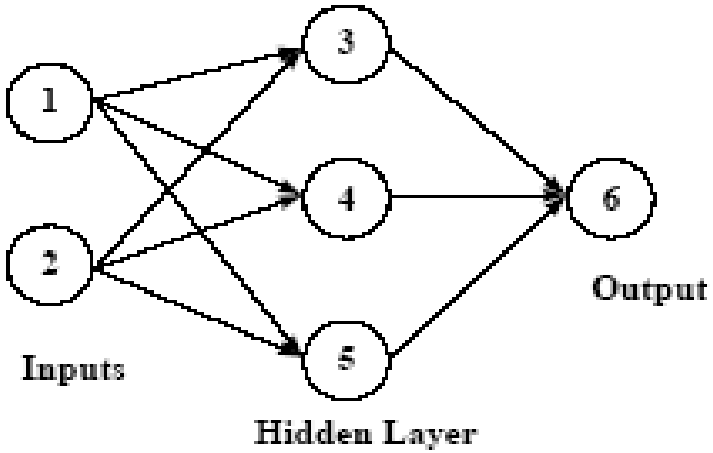


Невронни мрежи

- Multilayer Perceptron network of neurons
- Изчислява се вероятността на всяко възможно състояние на входните атрибути за сбъдване на всички възможни състояния на изходните атрибути
- Вероятностите се използват по-късно за прогнозиране
- За комплексни входни данни, при
  - наличие на голямо количество данни за обучение
  - липса на детерминистични правила
- Back-Propagated Delta Rule network
  - 3 нива неврони: входно, скрити, изходно
  - всеки неврон получава един или повече входове и създава един или повече идентични изходи
  - изходът е проста нелинейна функция на сумата на входовете
  - входовете се предават на междинното ниво и от там на изходите
  - няма връзка между невроните в едно ниво
- Един модел може да съдържа много мрежи

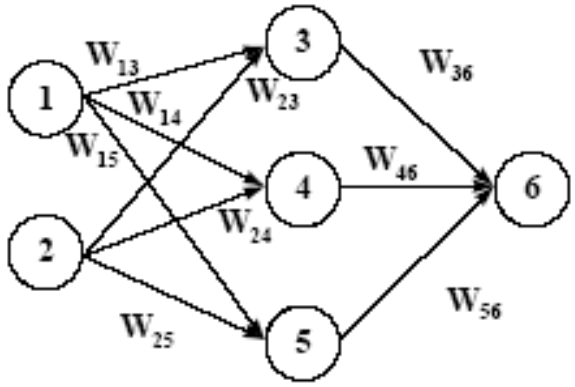
Невронни мрежи

- Топология
  - Входни неврони
    - всички стойности на входните атрибути (вкл. липсващи стойности)
  - Скрити неврони
    - получават стойности от входните и ги предават на изходните
  - Изходни неврони
    - състоянията на прогнозните променливи



Невронни мрежи

- Всеки вход има стойност, тегло, което определя значението му за скритите неврони
- Стойностите на теглата са неизвестни предварително и се изчисляват при обучение на модела
- Алгоритъм определя дали даден вход може да се използва за класифициране в определени случаи
- Стойността на входа се умножава по теглото
- Към всеки неврон е присвоена проста нелинейна функция, *activation function*, която задава значението му за нивото



Multimedia Mining

- Извличане на информация за свойства на мултимедийни обекти
- Обекти на търсене
  - образи
  - звук
  - видео
- Свойства на обектите
  - времеви
  - пространствени
- Методи
  - търсене по описание на обекта
  - търсене по съдържание, с или без описание

Multimedia Mining

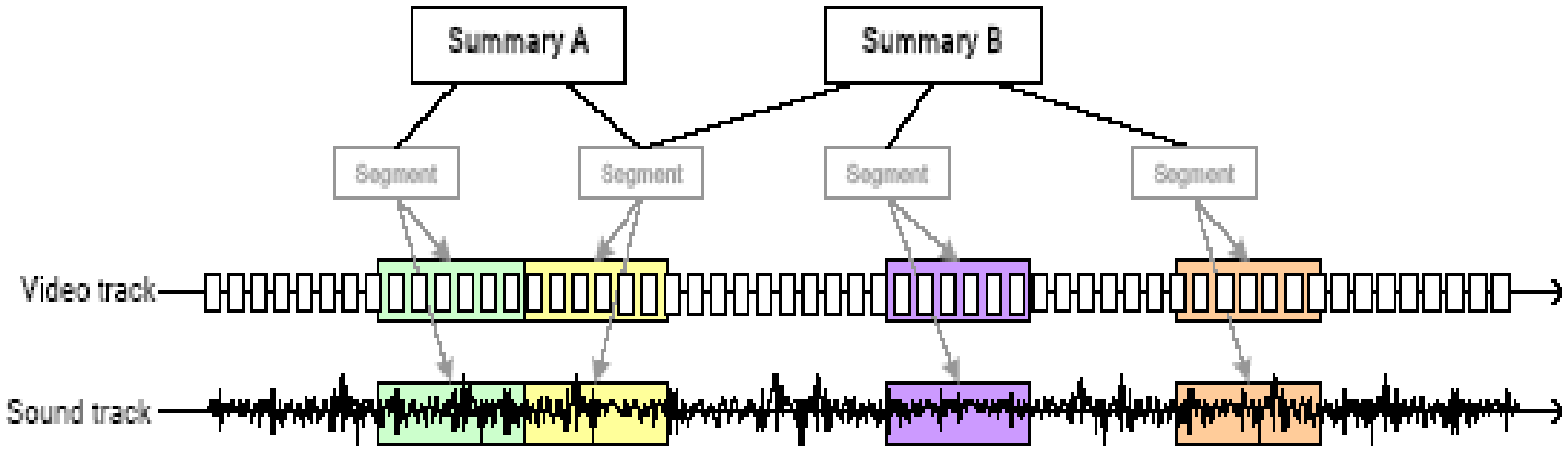
- Проблеми
  - хетерогенни статични и динамични мултимедийните обекти с различен произход и природа
  - обектите са носители на многократно повече информация от текстовите документи
  - многомерно пространство на свойствата на обектите
  - интерпретацията на тази информация е субективно определена
- Техники за търсене
  - обратна връзка за оценяване на приложимостта и самообучение на системата

Multimedia Mining

- Специализирани техники за търсене на
  - изображения
    - семантично представяне на изображения посредством структури (напр. граф от свързани изображения с оценка на свързаността)
  - звук
    - чрез описания
  - видео
    - създаване на домейни от пространствени обекти, които да се използват за видео разпознаване
    - анотиране и класификация чрез хомогенни текстури
    - създаване на пространствени асоциации между видео и семантични събития

Извличане на семантично знание

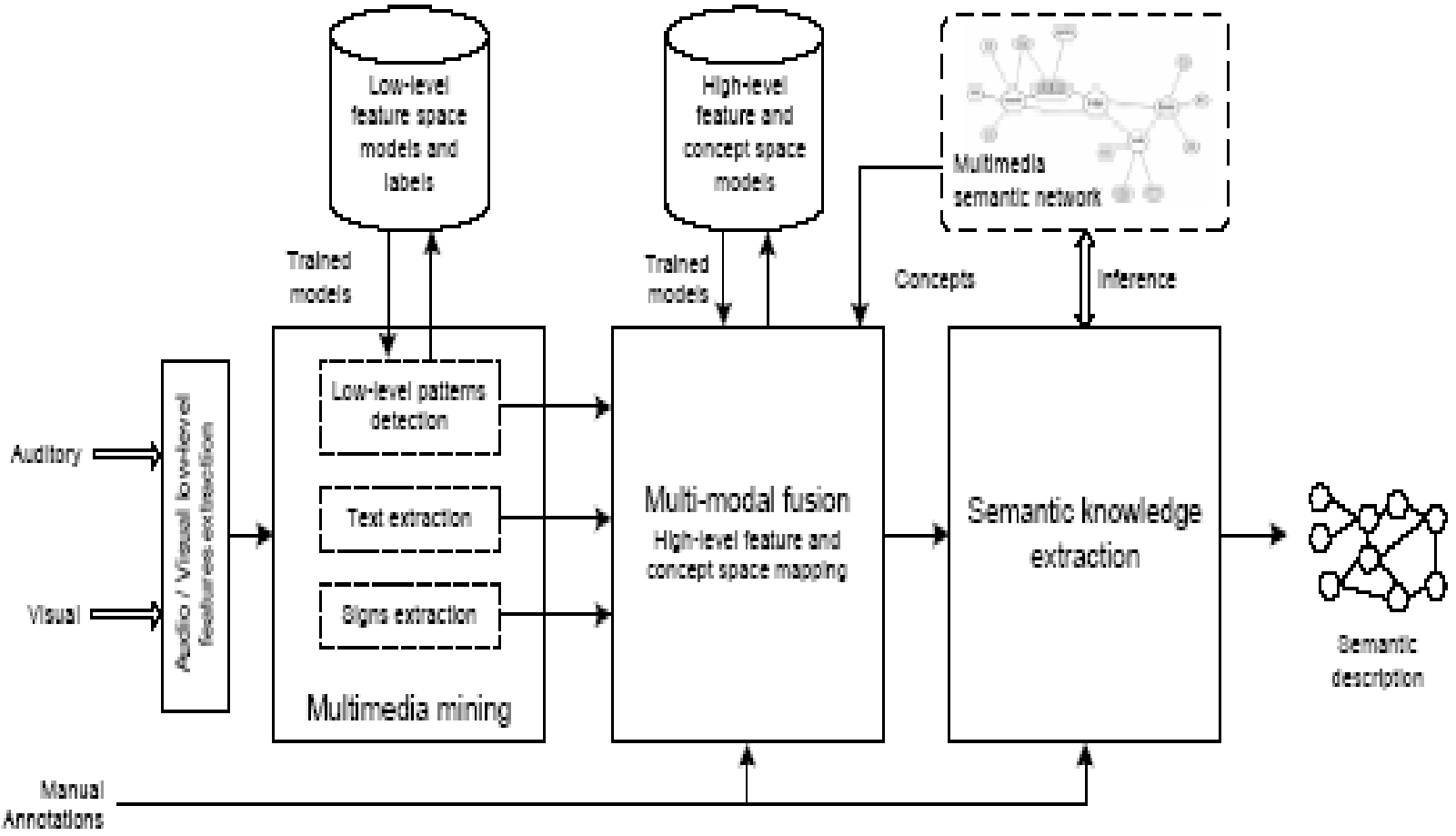
изборът на продължителност и обхват е персонален и съдържа семантично знание



Извличане на семантично знание

- Примери
  - дефиниране на семантични свойства като навън/вътре, гора, град, небе, вода, скала и т.н.
  - дефиниране на вероятностни асоциации между прости звукови и видео кадри
  - думи, асоциирани с цяло изображение и отделни райони от него и моделиране на вероятностно разпределение на районите и думите

Процедура


Low-level feature space models and labels
High-level feature and concept space models
Multimedia semantic network
Trained models
Trained models
Concepts
Inference
Auditory
Visual
Audio / Visual low-level feature extraction
Low-level patterns detection
Text extraction
Signs extraction
Multimedia mining
Multi-modal fusion
High-level feature and concept space mapping
Semantic knowledge extraction
Semantic description
Manual Annotations

Визуализация на информацията

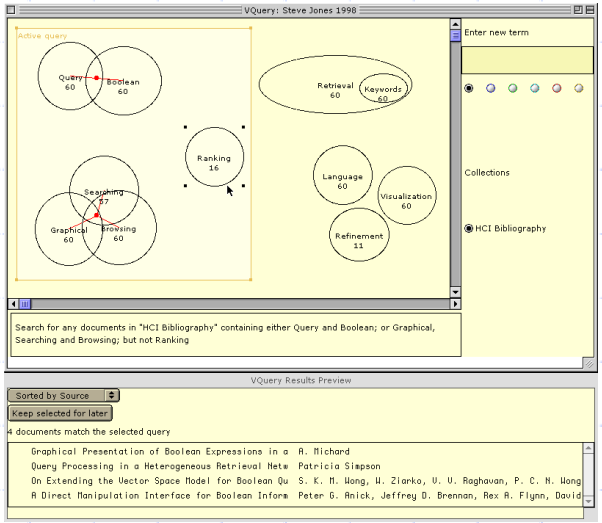
- Изключително динамично развиващо се направление
- Двумерна и тримерна визуализация на физични обекти и свойства по описанията им
- Динамична и интерактивна визуализация
- Визуализация на процеса и на резултатите

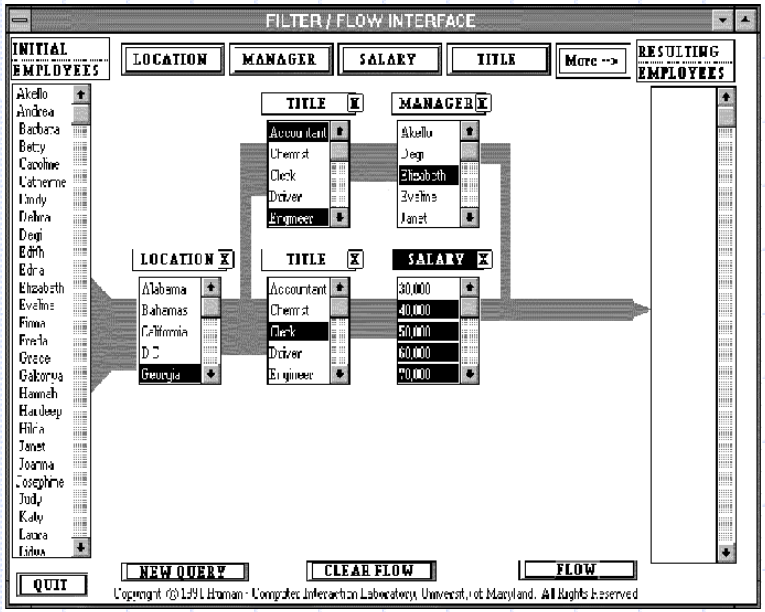
Техники за визуализация

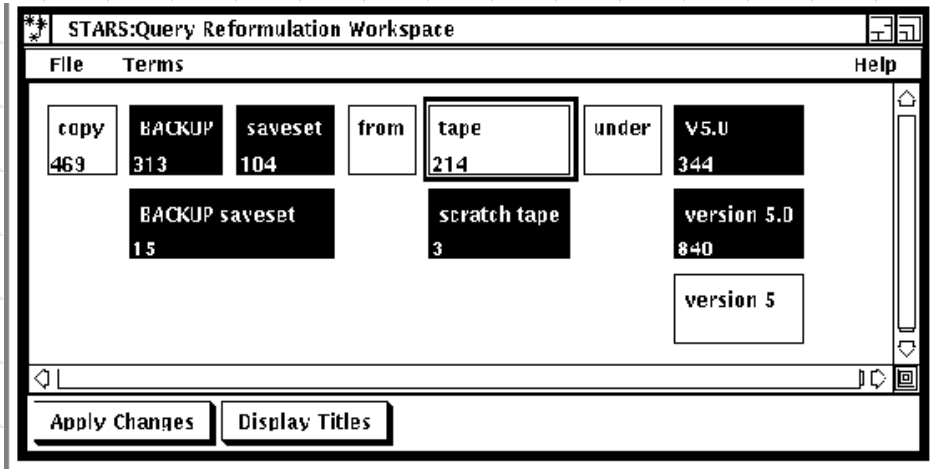
- **icons**
- **color highlighting**
- **brushing and linking**
  - свързване на два или повече изгледа на едни и същи данни
  - промяна в единия води до адекватна промяна в другия
  - използва се цветово кодиране
- **panning and zooming**
  - моделиране на изгледи от видеокамера
- **focus-plus-context**
  - акцентиране на части от данните
  - отделяне на съседите им
- **magic lenses**
  - трансформация на изображение на данни в резултат на визуална операция
  - създават се различни изгледи на изображението
- **animation**
  - за изобразяване на йерархии
- **overview plus details**
  - съчетано използване на техниките

Визуализация на процеса

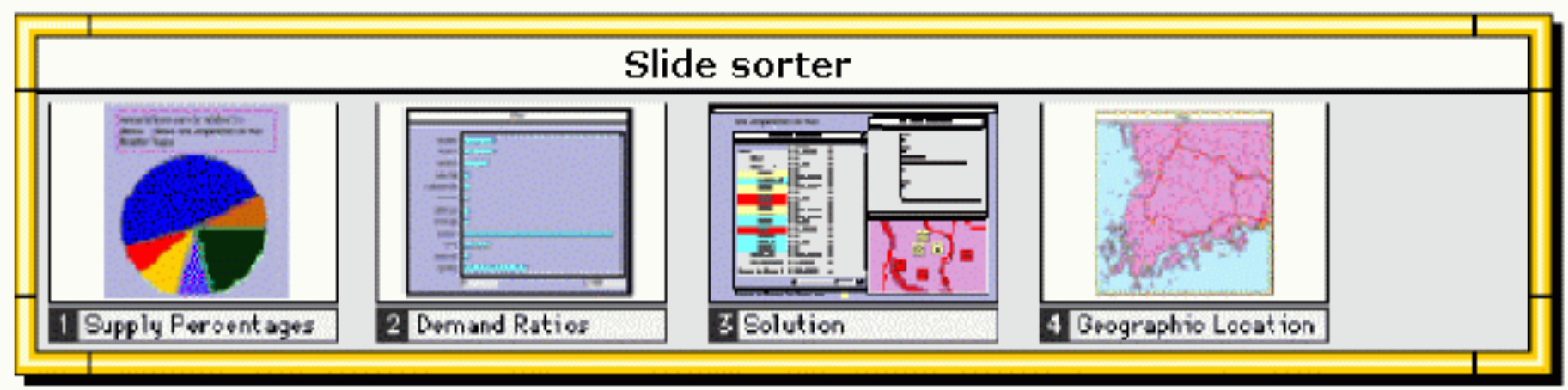
- Директна манипулация





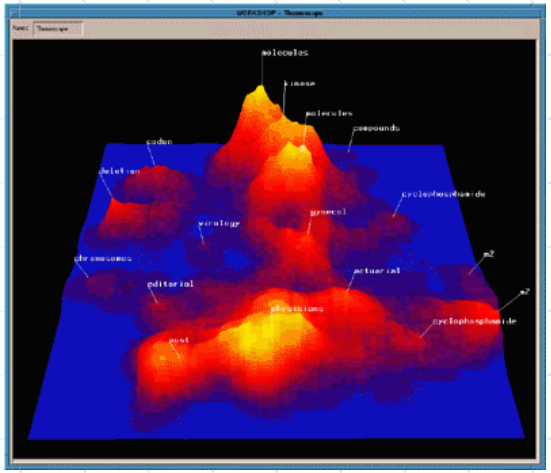


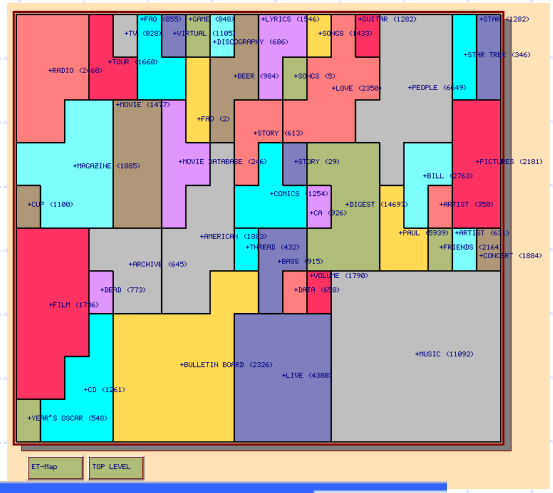
Визуализация на историята на процеса

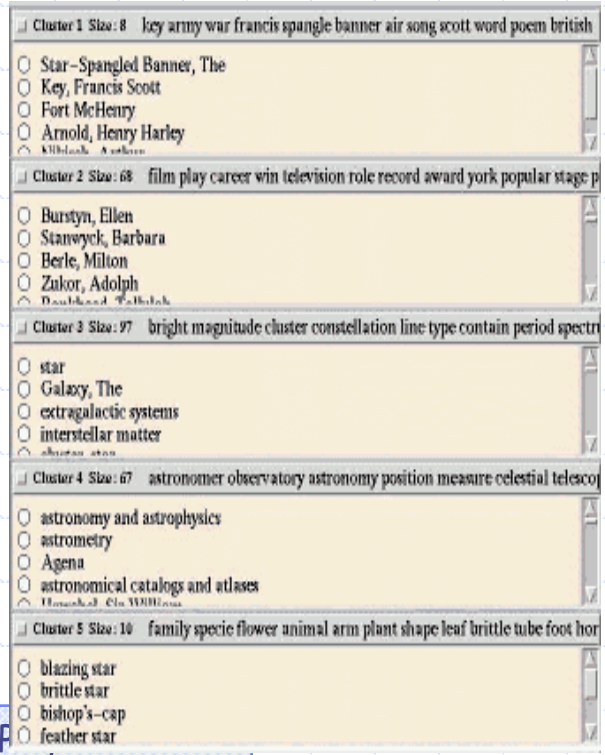


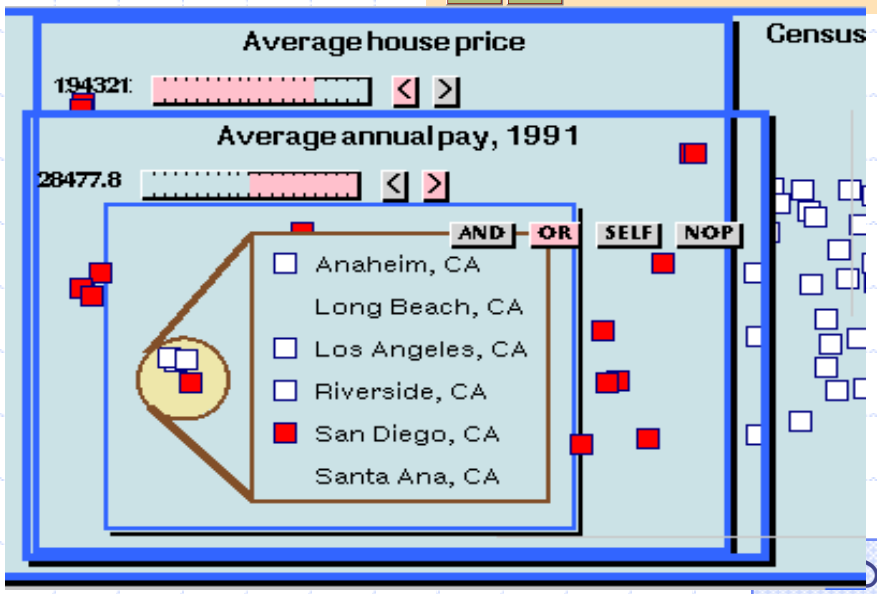
Визуализация на резултатите

- Карти
- Пространства
- Клъстери

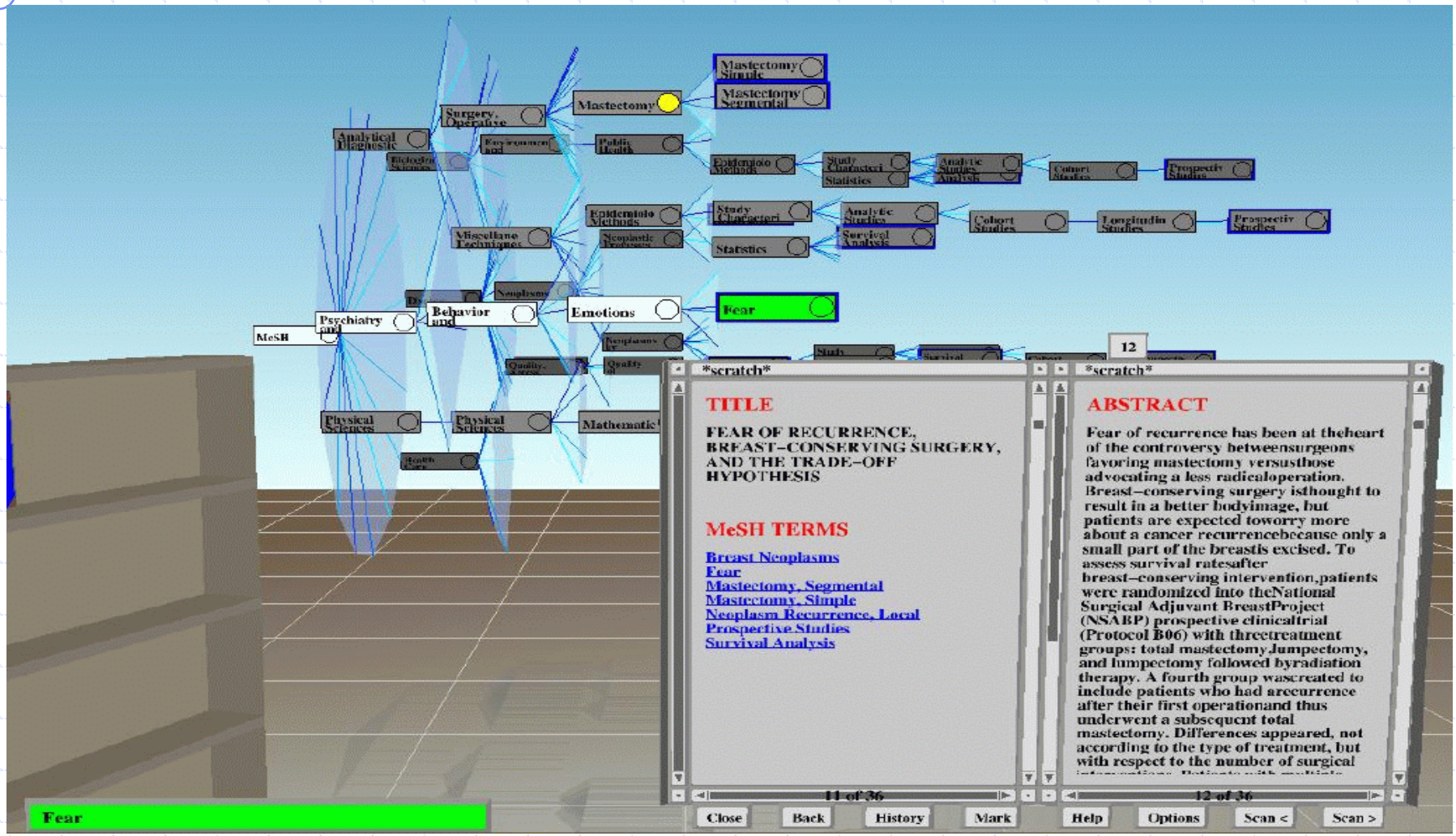




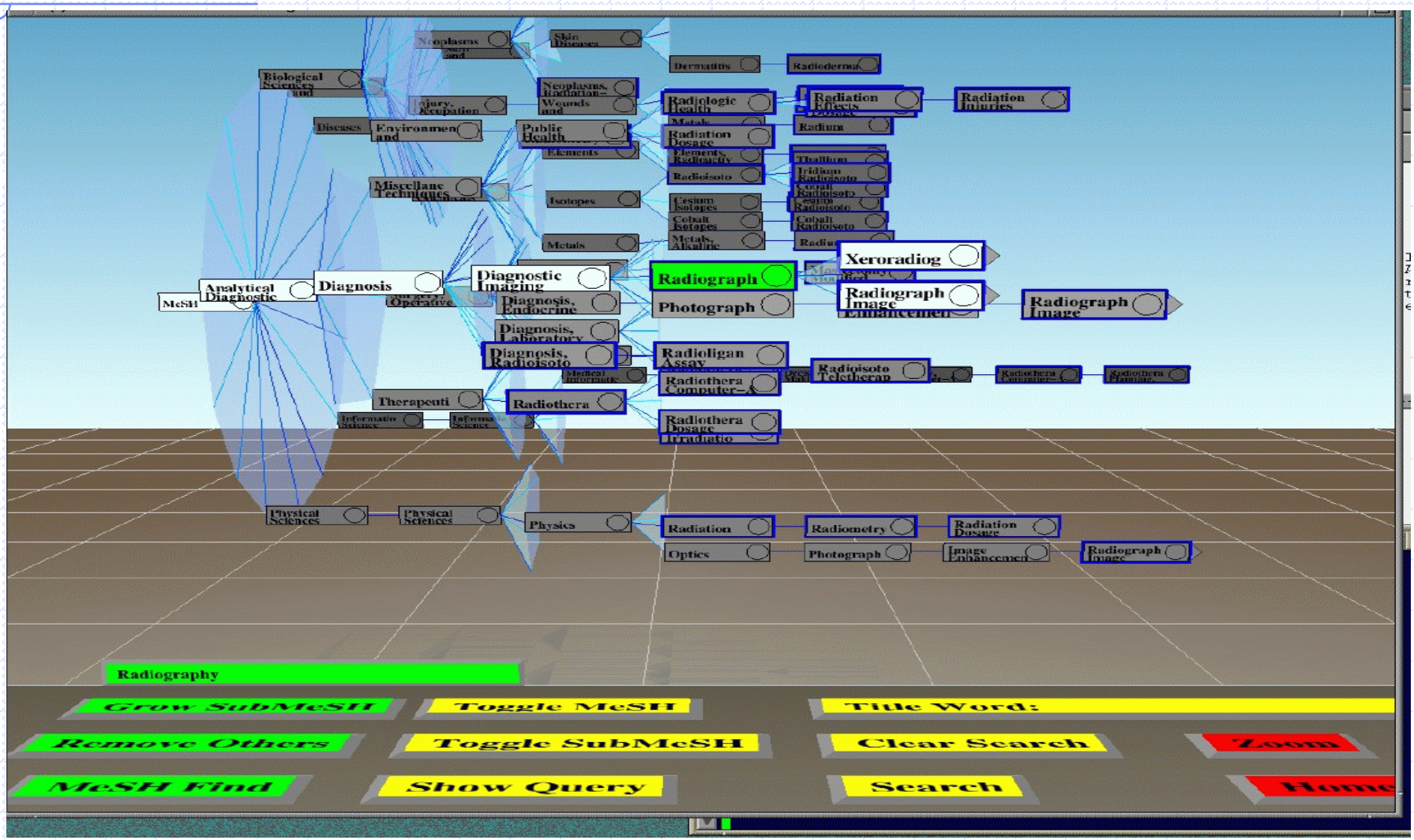




Интегриране на визуализацията



Интегриране на визуализацията



Тенденции

- Автоматично индексиране на маркирани документи
- Търсене в свободен, немаркиран текст
- Търсене в мултимедия и паралелни информационни потоци
- Синхронизация на времето за търсене от различни сървъри
- Визуализация на големи количества абстрактна информация
- Персонализация на търсенето и визуализацията
- Софтуерни агенти
- ...