Что было...

1. Введение

- 1.1. Цели курса:

- дать представление об общих задачах анализа текстов (= символьных последовательностей), возникающих в различных предметных областях и способах их решения;
- дать общий взгляд на уже известные алгоритмы , связанные с методом динамического программирования, и представить некоторые новые алгоритмы;
- дать навык формальной постановки задач анализа текстов;
- дать опыт анализа качества результатов алгоритмов анализа текстов на примере задачи восстановления скрытых состояний HMM;
- познакомить с примерами задач анализа текстов, специфическими для отдельных предметных областей (в основном, на примере биоинформатики)

Основные задачи

- сопоставление в целом (парное, множественное);

- определение количественной меры сходства последовательностей в целом;

- поиск общих мотивов (в двух и нескольких последовательностях); поиск в базах данных;

- поиск и выделение функционально значимых участков (заданных «паттернов»);

- разбиение последовательности на «однородные» участки;

- определение статистической значимости результатов сравнения и поиска.

При анализе данных требуется:

1. УМЕТЬ ПРИДУМЫВАТЬ не только алгоритмы, но и ФОРМАЛЬНЫЕ ПОСТАНОВКИ ЗАДАЧ.

2. ПРОВЕРЯТЬ, насколько ФОРМАЛЬНАЯ ПОСТАНОВКА ЗАДАЧИ соответствует СОДЕРЖАТЕЛЬНОЙ ПОСТАНОВКЕ.

1. Введение

1.2. Источники символьных последовательностей:

- тексты на естественных языках (художественная литература, деловая переписка, Интернет и т.п.);
- тексты на формальных языках (программы и т.п.);
- структуры биополимеров (ДНК, РНК, белки);
- цифровые данные (например, полученные при распознавании речи).

1. Введение

1.3. Общие для различных предметных областей задачи анализа последовательностей:

- сопоставление в целом (парное, множественное);

- определение количественной меры сходства последовательностей в целом;

- поиск общих мотивов (в двух и нескольких последовательностях); поиск в базах данных;

- поиск и выделение функционально значимых участков (заданных «паттернов»);

- разбиение последовательности на «однородные» участки;

- определение статистической значимости результатов сравнения и поиска.

2. Динамическое программирование на графах

- «Задача Беллмана»: поиск минимального пути в ориентированном ациклическом графе с весами на ребрах (ОАВГ). Веса – действительные числа. Уравнение Беллмана. Время работы ~ $N_{ребер}$. Память ~ $N_{вершин}$.
- Вычисление полной обобщенной статистической суммы (ОСС) для данного ОАВГ («Задача Больцмана»). Связь с задачей Беллмана. Частные ОСС для отдельных вершин. Вычисление всех ЧОСС
- ОАВГ с весами в полукольце. “Соединительная” операция (умножение) и “целевая” операция (сложение). Вычисление ОСС для ОАВГ при известном топологическом порядке на вершинах. Некоммутативность умножения и коммутативность сложения. Примеры.
- Накопление результатов. Связь между алгоритмами Беллмана и Дейкстры. Алгоритм A*. *Оптимистический поиск.*
- *Векторные веса и Парето-множества.*

2. Dynamic programming on graphs.

- Search for minimal path in the weighted directed acyclic graph (WDAG). Bellman equation.
- Calculation of partition function. Relation to minimal path problem.
- WDAGs with weights in a semiring. “Joining” operation (multiplication) and “target” operation (addition). Calculation of partition function for a WDAG with known topological ordering of vertexes. (Non)commutativity of addition and multiplication.
- Special partition functions and their calculation.
- Algorithm with accumulation of results. Comparison with Bellman’s and Dijkstra’s algorithms. *Optimistic search.
- Complexities.

* * Vector weights

3а. Глобальное выравнивание

3.1.Неформальная постановка задачи. Связь с эволюцией.

3.2. Вес выравнивания. Матрица весов замен. Инделы и их веса (штрафы).

3.3. Алгоритм для посимвольных штрафов за инделы. Сведение к задаче Беллмана. Граф выравнивания. Матрица выравнивания. Сложность алгоритма.

3.4. Другие постановки задач и алгоритмы.

- аффинные штрафы за инделы; штрафы за инделы общего вида;

- выравнивание в полосе;

- значимость отдельных сопоставлений, связь с частичными ОСС;

- особые (нулевые) штрафы за концевые инделы;

- *штрафы за разрезы;*

- *оптимистический алгоритм для случая диагональной матрицы замен (совпал – не совпал) и посимвольных штрафов за инделы.*

3. Global pairwise sequence alignment

- 3.1. Informal problem statement. Role of evolution hypothesis.
- 3.2. Alignment weight. Match weight. Indels (gaps) and indel penalty.
- 3.3. Algorithm for 1-symbol indel penalty. Reduction to graph problem.
- 3.4. Other problem statements: affine gap penalties; general gap penalty function;
- matching significance based on special partition fubction. Reduction to graph
- problems. Role of gap penalties for ending fragments.
- 3.5. Role of etalon (“golden standard”) alignments. Quality of algorithmic alignment (accuracy and confidence). Verification of problem statement. Recourses for improvement of alignment quality, usage of extra-sequential information.

3б. Глобальное выравнивание

3.5. Качество алгоритмического выравнивания. Точность (accuracy) и достоверность (confidence). Роль эталонных выравниваний (“golden standard”). Источники эталонных выравниваний: выравнивания реальных данных, полученные с помощью экспертизы и/или дополнительных данных, и модельные выравнивания. Валидация постановки задачи.

3.6. Ресурсы для улучшения качества выравнивания: использование дополнительной информации.

- Examples of algorithmic problems that cannot be reduced to graphs.
- Hypergrahs, hyperedges and hyperpathes. Weighted directed hypergraphs. Acyclic hypergraphs.
- Semirings. Bellman’s algorithm for calculation of partition function of a given weighted directed acyclic hypergraph (WDAHG) with known topological ordering of vertexes. Complexity.
- Calculation of special partition functions. Additional restrictions: commutativity of multiplication; strong aciclicity. Complexity.
- Analogs of algorithm with accumulation of results and Dijkstra’s algorithms.

4.1. Алгоритмические задачи, несводимые к задаче поиска минимального пути в ОАВГ.

4.2. Ориентированные гипреграфы. Гиперребра, гиперпути, веса гиперграфы. Ациклические (строго ациклические) гиперграфы. Веса из произвольного полукольца.

4.3. Вычисление полной ОСС для ориентированного ациклического гиперграфа с весами на гиперребрах (ОАВГГ) при известном топологическом порядке на вершинах гиперграфа. Время работы ~ суммароное к-во концов гиперребер. Память ~ к-во вершин.

4.4. Частные ОСС для отдельных вершин. Вычисление частных ОСС для всех вершин при доп. ограничениях: строгая ацикличность, коммутативность умножения. Сложность алгоритма.

4.5. Аналоги алгоритма с накоплением результата и

- Examples of algorithmic problems that cannot be reduced to graphs.
- Hypergrahs, hyperedges and hyperpathes. Weighted directed hypergraphs. Acyclic hypergraphs.
- Semirings. Bellman’s algorithm for calculation of partition function of a given weighted directed acyclic hypergraph (WDAHG) with known topological ordering of vertexes. Complexity.
- Calculation of special partition functions. Additional restrictions: commutativity of multiplication; strong aciclicity. Complexity.
- Analogs of algorithm with accumulation of results and Dijkstra’s algorithms.

- Examples of probability models: Bernoulli model, Markov model, Hidden Markov model (HMM).
- HMMs and graphs. Involute of HMM graph. Trajectory graph of an HMM.
- Sequence of states corresponding to a given sequence. Viterbi algorithm and forward-backward algorithm.
- HMMs and dynamic programming. Viterbi algorithm and optimal path in trajectory graph. Forward-backward algorithm and special partition function on trajectory graph.
- How to estimate quality of state assignment made by an algorithm?
- Estimation of parameters of a probability model. Problem statement.
- EM-algorithm.
- Baum-Welch algorithm.

- HMMs and finite automata. Computation of probability of set of sequences. General construction and examples.
- Seed sensitivity.
- Statistical significance of cluster of pattern entries.
- Sequence segmentation. Example: finding coding regions in DNA sequences. Types of HMMs describing procariotic genome. Problem of boundaries.

- Multiple sequence alignment. Dynamic programming without and with restriction on gap size.
- Profiles (Position Specific Weight Matrixes, PSWMs) . Alignment of a sequence and a profile; alignment of profiles. Progressive alignment.
- Pattern recognition. Description of patterns with PSWMs and consensus words.

7.1. Sequence segmentation. Example: finding coding regions in DNA sequences. Types of HMMs describing procariotic genome. Problem of boundaries.

СДЕЛАТЬ:

7.2. Задача о разладке – статистические подходы.

- Statistical significance of cluster of pattern entries.

- СДЕЛАТЬ:

****** ТеорияКАРЛИНА-АЛЬТШУЛЯ, Мотта и BLAST.**

10.1. Статистическая значимость кластера сайтов. Графы перекрытий.

10.2. Динамическое программирование на решетке. Использование деревьев поиска. Мажорирование и «хозяева областей». *Списки кандидатов.*