王 冰

E-mail：wangbing@ict.ac.cn Phone：(+86)156-3885-9198 Blog: http://blog.csdn.net/zzucaicai

教育背景

中国科学院计算技术研究所（保送 前 2%） 2014.09 - 2017.07

工学硕士，前瞻研究实验室 研究方向：基因组序列拼接算法研究

郑州大学 2010.09 - 2014.07

工学学士，计算机科学与技术专业 专业排名：**2 / 89**

项目经历

基因组序列拼接算法的研究及拼接软件 ARCS 的实现 2015.08 - 今

核心开发人员 **中科院计算所**

- 简介：从百万量级的短序列片段中恢复出原始 DNA 序列。该问题可形式化为求解序列 overlap 图的汉密尔顿回路(路径）问题，图中节点表示短序列片段，边表示短序列片段之间的重叠区域。由于汉密尔顿回路（路径）问题是 NP 完全问题，因此将其转化为求解 De bruijn 图的超欧拉回路（路径）问题。
- 职责：
  - 利用 hash 策略降低节点的内存需求并将建图的时间复杂度由 $O(N^2)$ 降低到 $O(kN)$。
  - 基于片段之间的距离信息确定片段在 DNA 上的相对位置。利用距离矩阵本身的稀疏性，将问题转化成求解矩阵 L1 范数最小化问题，使用**线性规划**模型求解。
  - 设计判定重复序列（序列拼接中的主要挑战）算法。利用改进的**混合高斯模型**和 BIC 准则对距离信息聚类，采用 **EM 算法**迭代求解。利用聚类中心个数判定重复序列。
  - 利用 boost 库、单件注册列表等技术使软件易读、易维护、可扩展。利用**多线程**优化大批量数据处理。
- 成果：
  - 完成 ARCS 软件代码编写并开源(https://github.com/bigict/ARCS)。与北京基因组所合作将 ARCS 投入使用。
  - ARCS 测试结果(准确度相同时的拼接长度)优于目前主流软件约 **10%**。
  - 相较于传统局部拼接策略，ARCS 得到全局最优的拼接结果。
- **关键字**：混合高斯模型，EM 算法，线性规划，c++，多线程，boost

程序猿信息检索系统 2014.12 - 2015.01

核心开发人员 **中科院计算所**

- 简介：系统抓取 CSDN 博客网页为程序员提供搜索服务。实现关键字查询，文本聚类，前缀查询提醒等功能。
- 职责：
  - 对用户查询，返回相关性高的文档。利用**向量空间模型**和 **TF-IDF** 表示文档，并过滤 TF-IDF 低的词项。对词项文档矩阵做隐性语分解（**SVD 分解**），提高查询的**召回率**。
  - 实现不同主题文档分类展现。利用 **Kmeans** 和**层次聚类**方法对文档聚类，从而实现文档分类展现。
  - 实现查询自动补全功能。利用 **Trie 树**实现前缀匹配及模糊匹配，以提示用户可能的输入内容。

实习经历

2013.10 - 2013.12 **金山云** **分布式文件系统测试及性能优化**

- 职责：安装配置 **MooseFS**，测试各个参数对性能的影响，查找性能瓶颈。用汇编语言改写 crc 校验部分代码。
- 收获：crc 校验速度提升 **20%**，对分布式文件系统有整体了解，熟练使用常用 linux 命令。

个人技能

- 编程能力：熟悉 C++，熟悉面向对象基本思想及常用设计模式，了解常用 boost 库；了解 Java，Python。
- 算法能力：良好的数据结构和算法基础，曾两次担任国科大《算法分析与设计（卜东波）》课助教。
- 机器学习：了解常用机器学习算法（LR、GBDT、随机森林、隐马模型）。

获奖情况及其它

- 2012 37 届 ACM-ICPC 国际大学生程序设计竞赛亚洲区 金华站 **银奖**
- 2013 38 届 ACM-ICPC 国际大学生程序设计竞赛亚洲区 成都站，长沙站 **铜奖**
- 2015 中国大学生程序设计竞赛 **银奖**
- 2012 国家奖学金（**前 1%**）
- 2010/2011/2012 郑州大学一等奖学金（**三次**）（**前 5%**）

Bing Wang

E-mail：wangbing@ict.ac.cn Phone：(+86)156-3885-9198 Blog: http://blog.csdn.net/zzucaicai

Education

University of Chinese Academy of Sciences Sep. 2014 - Present

Master's degree, Institute of Computing Technology Research area: Genome Assembly

Zhengzhou University Sep. 2010 - Jun. 2014

Bachelor's degree, College of Information and Technology GPA: 3.80/4.0 Rank: **2/89**

Experience

Genome Assembler – ARCS(Assemble short-read via combinatorial optimization in scaffolding**)** Aug. 2015 - Present

Core Developer ICT

- **Brief:**
  - Given Millions of medium-sized DNA fragments, we need to tie those fragments together to obtain the original DNA sequence.
  - This problem can be formalized as a Hamiltonian Circle problem in graph, in which each fragment is represented by a node and overlap between fragments represented by a directed edge.
  - Due to the computational difficulties of Hamiltonian Circle problem, it is transformed into an Euler Circle problem in de Bruijn Graph.
- **Responsibilities:**
  - Using hash strategy to decrease the memory demand of nodes and reduce the time complexity of Building the graph.
  - Responsible for determining the relative order of fragments. Linear Programming Model is used to obtain the positions of fragments.
  - Responsible for finding out the repeated fragments. **Gaussian Mixture Model (GMM)** and **Bayesian Information Criterion (BIC)** is used to cluster the distance information to determine repeats. **EM** iterator is used to estimate parameters.
  - **The boost library** and **registry of singleton** are used to improve the readability, maintainability and scalability of the software. Multi-threading is used to optimize bulk data processing.
- **Achievements:**
  - The open source software ARCS is completed and available at https://github.com/bigict/ARCS. And ARCS is now used for research in collaboration with Beijing Institute of Genomics.
  - Compared with mainstream assemblers, ARCS gets a improvement about **10%.**
- **Key Words:** GMM，EM, Linear Programming，C++，Multi-threading，Boost

Information Retrieval System for Programmers Dec. 2014 - Jan. 2015

Core Developer ICT

- **Brief：** A system that grabbing the pages of CSDN blogs provides searching services for programmers.
- **Responsibilities:**
  - Responsible for returning appropriate documents for a query. **Vector space model** and TF-IDF is used to represent documents. **SVD** matrix decomposition is used to build latent semantic indexing to improve the recall rates.
  - Responsible for clustering the returning documents based on different themes. The clustering strategy **K-means** and **hierarchical** clustering is used to display the documents clustered by themes.
  - Responsible for implementing query automatic completion. **Trie tree** is used for prefix matching and fuzzy matching to prompt users about the possible query sentences.

Internship

2013.10 - 2013.12 **Kingsoft Cloud** **Testing and optimizing distributed file system**

- **Responsibilities**: Responsible for the installation and configuration of **MooseFS**，testing the influence of each parameter，and finding out the bottleneck of the performance. Rewriting the code of CRC in assembly language.
- **Achievements**: The running speed of CRC increased by **20%**.

Technical Strengths

- Familiar with C++, Object-Oriented programming and design patterns. Know frequently used boost library. Know Java, python.
- Familiar with basic algorithms and data structures.
- Know common machine learning algorithms(Logistic Regression, GBDT, Random Forest, Hidden Markov Model).

Rewards

- 2012 37th ACM-ICPC Asia Jinhua regional contest **silver medal**
- 2013 38th ACM-ICPC Asia Chengdu and Changsha regional contest **bronze medal**
- 2015 Chinese collegiate Programming contest **silver medal**
- 2012 National Scholarship (**Top 1%**)
- 2010/2011/2012 First class award Scholarship (**Top 5%**)
- 2015 Excellent Student in UCAS