Reference

OJ:

http://lintcode.com/

https://oj.leetcode.com/

Website:

http://www.ninechapter.com/

(new version coming soon)

http://www.geeksforgeeks.org/

http://www.glassdoor.com/

http://www.careercup.com/

Book:

<<Cracking The Coding Interview>>

<<编程之美>>

Wechat Subscription:

9. Resume & Big Data

九章算法IT求职面试培训 第9讲

www.ninechapter.com

Resume

how to get the chance of phone screen

What you shouldn’t present in your resume

Something not related to CS (e.g. Association)

Too long

- New Grad - 1 Page
- Junior Experienced - 1-2 Pages
- Senior Experienced - 2-3 Pages

Bad formatted. (font style, font size, alignment)

English Version + Chinese Version

NO doc/docx (MUST BE PDF)

Low GPA

Proficient in C++

What does the HR focus on?

Where are you from (Education)

What offers you got before (Intern/Experience)

What you did before (Projects)

What can you do (Skills)

What kind of people you are (Awards, Self-Evaluation)

- Self-Drive
- Communicative / Collaborative
- Smart / Fast learning

Project / Intern / Experience

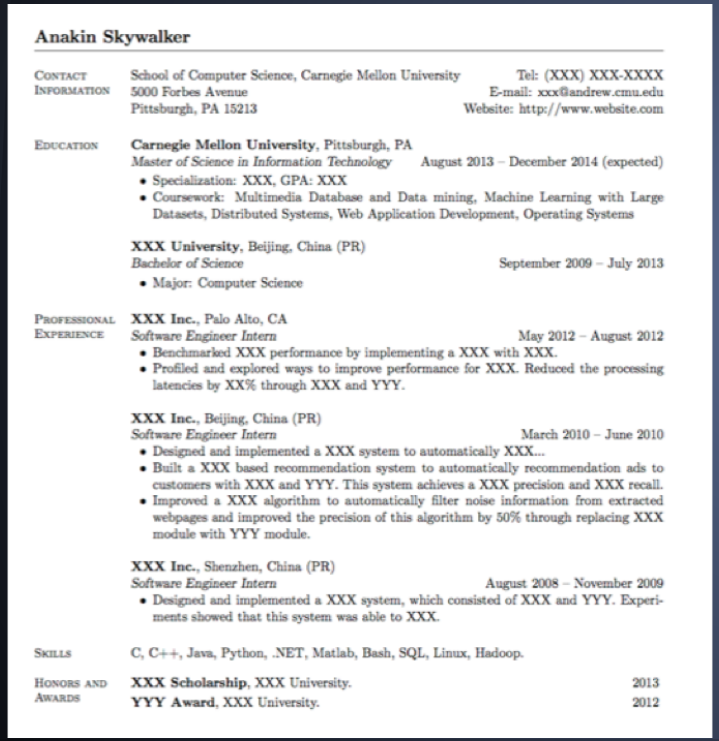
What is this project?

What I did in this project?

What is my impact? (Mostly ignored)

- Achievement (speed up 5%, precision up 5%)
- User (How many users are using your product)

Examples



Anakin Skywalker

Contact Information	School of Computer Science, Carnegie Mellon University 5000 Forbes Avenue Pittsburgh, PA 15213	Tel: (XXX) XXX-XXXX E-mail: xxx@andrew.cmu.edu Website: http://www.website.com

Education

Carnegie Mellon University, Pittsburgh, PA
Master of Science in Information Technology — August 2013 – December 2014 (expected)
- Specialization: XXX, GPA: XXX
- Coursework: Multimedia Database and Data mining, Machine Learning with Large Datasets, Distributed Systems, Web Application Development, Operating Systems

XXX University, Beijing, China (PR)
Bachelor of Science — September 2009 – July 2013
- Major: Computer Science

Professional Experience

XXX Inc., Palo Alto, CA
Software Engineer Intern — May 2012 – August 2012
- Benchmarked XXX performance by implementing a XXX with XXX.
- Profiled and explored ways to improve performance for XXX. Reduced the processing latencies by XX% through XXX and YYY.

XXX Inc., Beijing, China (PR)
Software Engineer Intern — March 2010 – June 2010
- Designed and implemented a XXX system to automatically XXX...
- Built a XXX based recommendation system to automatically recommendation ads to customers with XXX and YYY. This system achieves a XXX precision and XXX recall.
- Improved a XXX algorithm to automatically filter noise information from extracted webpages and improved the precision of this algorithm by 50% through replacing XXX module with YYY module.

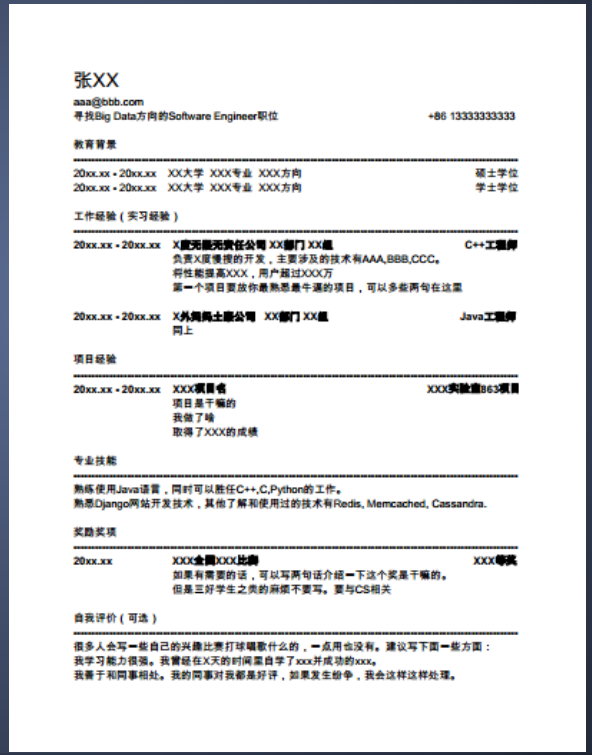
XXX Inc., Shenzhen, China (PR)
Software Engineer Intern — August 2008 – November 2009
- Designed and implemented a XXX system, which consisted of XXX and YYY. Experiments showed that this system was able to XXX.

Skills

C, C++, Java, Python, .NET, Matlab, Bash, SQL, Linux, Hadoop.

Honors and Awards

XXX Scholarship, XXX University. 2013
YYY Award, XXX University. 2012



张XX

aaa@bbb.com
寻找Big Data方向的Software Engineer职位 +86 13333333333

教育背景

20xx.xx - 20xx.xx XX大学 XXX专业 XXX方向 硕士学位
20xx.xx - 20xx.xx XX大学 XXX专业 XXX方向 学士学位

工作经验（实习经验）

20xx.xx - 20xx.xx **X度无限无责任公司 XX部门 XX组** **C++工程师**
负责X度慢慢的开发，主要涉及的技术有AAA,BBB,CCC。
将性能提高XXX，用户超过XXX万
第一个项目要放你最熟悉最牛逼的项目，可以多些两句在这里

20xx.xx - 20xx.xx **X外网码土豪公司 XX部门 XX组** **Java工程师**
同上

项目经验

20xx.xx - 20xx.xx **XXX项目名** **XXX实验室863项目**
项目是干嘛的
我做了啥
取得了XXX的成绩

专业技能

熟练使用Java语言，同时可以胜任C++,C,Python的工作。
熟悉Django网站开发技术，其他了解和使用过的技术有Redis, Memcached, Cassandra.

奖励奖项

20xx.xx **XXX全国XXX比赛** **XXX等奖**
如果有需要的话，可以写两句话介绍一下这个奖是干嘛的。
但是三好学生之类的麻烦不要写。要与CS相关

自我评价（可选）

很多人会写一些自己的兴趣比赛打球唱歌什么的，一点用也没有。建议写下面一些方面：
我学习能力很强。我曾经在X天的时间里自学了xxx并成功的xxx。
我善于和同事相处。我的同事对我都是好评，如果发生纷争，我会这样这样处理。

Big Data

A wave of data is coming

Most frequent IP address

Find the most frequent IP address in web log.
The log file size may over 100G

Naive method

Use HashMap to count each ip

Time: O(N * READ)
Memory: O(N')

Optimize 1

Only 2^32 distinct IP address → 4G, use 4 (sizeof int) * 4G = 16G memory.

What if I have 100M memory?

Optimize 2

1. Split 100G data into 1000 files, route function is “hash(ip) % 1000”.
2. Go through each file again, get most frequent ip for each file.

Time: O(2 * N * READ + N * WRITE)

Memory: O(N’ / 1000)

Disadvantage: Need more disk space, too much write.

Optimize 3

Lossy Counting, Sticky Sampling, Space Saving, Count Sketch, Efficient Count, hCount …

Time: O(N * READ)

Memory: O(1) Amazing!

Disadvantage: Probability Algorithm, may lose accuracy.

Top K frequent IP addresses

Find the top K frequent IP addresses in web log.
The log file size may over 100G

Hash + Heap

Use a min-heap, keep the top k frequent items in heap, every time you find an item occurs more than the smallest item in the heap but not occurs in heap, replace the smallest one.

Time: O(N * READ + N * log(K))

Memory: O(N')

Optimize

Split into multiple files

- According to hash(ip) % 1000

Probability Algorithm

- hCount/hCount*
- → h[hashfunc(key) % MEMORY]++;

Bit Map

010101010
Membership (YES)
Counter (NO)

Bloom Filter

“I am a Hash Map”
Membership (YES)
Counter (YES)

Bloom Filter

How it works?

How it saves memory?

False Positive vs False Negative

Word Search in File System

Given a file system which may contains thousands of files on disk, 1K~100M per file. Given n words. For each word, find all files include the word. You have only 100K memory and couldn’t use any disk storage.

Trie

“I am Tree”

Top K frequent IP addresses in N servers.

On each server, we have 100G log data.

Most K frequent ip in N servers

Algorithm 1: Get most K frequent ip addresses for each server, merge N*K ip addresses together to get the global top K.

May have problems if the real top k not included in the N*K IP addresses.

Most K frequent ip in N servers

Algorithm 2: Get most 100K frequent ip addresses for each server, merge 100*N*K ip addresses together to get the global top K.

May have problems if the real top k not included in the 100*N*K IP addresses.

Cost more memory and CPU.

Most K frequent ip in N servers

Algorithm 3: Relocate each IP address by hashcode, make sure the same IP address store on the same server.

Huge cost of relocation!

Find common URLs in two large URL files.

5G URLs per file, 64bytes for each url. 4G memory.

Common Urls

Algorithm 1: Split file A into 1000 files, urls in each file has the same shard id - hash(url) % 1000. Import each file to memory one by one, use hash map to get the common urls.

Algorithm 2: Bloom Filter.

Conclusion

Do not have enough memory?

→ Use disk

→ Lose accuracy

Hash, Heap, Bloom Filter, Trie, BitMap

Q & A

感谢您坚持完成了九章算法课程的学习
祝找到理想工作！