ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Lê Đinh Hợp

Thuật toán đánh chỉ mục ngược với MapReduce và ứng dụng trong việc đánh giá ý kiến của học sinh Hòa Bình trên mạng xã hội

Chuyên ngành: **Khoa học máy tính**

Mã số : **60 48 01 01**

Người hướng dẫn khoa học: PGS TS Đỗ Trung Tuấn

Thái Nguyên, 12 - 2016

Lời cam đoan

Tôi xin cam đoan:

Những kết quả nghiên cứu được trình bày trong luận văn là hoàn toàn trung thực, của tôi, không vi phạm bất cứ điều gì trong luật sở hữu trí tuệ và pháp luật Việt Nam. Nếu sai, tôi hoàn toàn chịu trách nhiệm trước pháp luật.

TÁC GIẢ LUẬN VĂN

Lê Đinh Hợp

Lời cám ơn

Tôi xin chân thành cảm ơn Trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên đã tạo điều kiện thuận lợi cho tôi hoàn thành khóa học này.

Tôi xin chân thành cảm ơn các Thầy Cô giáo – Các nhà khoa học đã trực tiếp giảng dạy truyền đạt những kiến thức chuyên ngành Khoa học máy tính cho tôi trong những tháng năm học tập tại trường.

Đặc biệt tôi xin bày tỏ lòng biết ơn chân thành và sâu sắc tới PGS TS Đỗ Trung Tuấn đã tận tình hướng dẫn, dìu dắt và chỉ bảo cho tôi những kiến thức về chuyên môn thiết thực và những chỉ dẫn khoa học quý báu để tôi hoàn thành bản luận văn này.

Luận văn này còn nhiều thiếu sót, rất mong được các thầy cô giáo trong hội đồng chấm luận văn xem xét, góp ý kiến để luận văn được hoàn thiện hơn.

Tôi xin chân thành cảm ơn!

Thái Nguyên, tháng 12 năm 2016

Mục lục

Danh sách các từ viết tắt

CNTT	Công nghệ Thông tin
HDFS	Hadoop Distributed File System
URL	Uniform Resource Locator
HTML	HyperText Markup Language
LISP	LISt Processing
ML	Markup Language
HPC	High-Performance Computing
NAS	Network-Attach Storage
SAN	Storage Area Network
GFS	Google File System
SPOF	Single Point Of Failure
SNN	Secondary NameNode
APW	Associated Press Wordstream
REST	Representational State Transfer
PRAM	Parallel Random Access Machine
BSP	Bulk Synchronous Parallel

Danh mục các hình vẽ, bảng biểu

Chương mở đầu

Đặt vấn đề

Trong thời đại hiện nay, công nghệ thông tin được ứng dụng tại mọi lĩnh vực trong cuộc sống, với một hệ thống máy tính người ta có thể làm được rất nhiều công việc, tiết kiệm được thời gian, công sức và tiền bạc. Với sự phát triển vượt bậc của Internet hiện nay, lượng thông tin ngày càng nhiều, sự tăng trưởng có thể nói là được tính bằng cấp số nhân, theo một nghiên cứu thì cứ khoảng 5 năm thì lượng tri thức của nhân loại sẽ tăng gấp đôi, với lượng thông tin đồ sộ trên mạng hiện nay thì việc tìm kiếm và khai thác thông tin là một công việc hết sức quan trọng, mang lại nhiều lợi ích về khoa học và kinh tế.

Cùng với sự ra đời của Internet, sự xuất hiện và phát triển không ngừng của lĩnh vực thương mại điện tử, các lĩnh vực nghiên cứu xã hội khiến cho việc xúc tiến các hoạt động kinh doanh hoặc nghiên cứu, quảng bá sản phẩm dịch vụ diễn ra trên khắp các kênh thông tin xã hội, đặc biêt là trên Internet.

Như chúng ta đã biết ngày nay mọi thông tin đều được đưa lên các trang mạng xã hội dưới dạng các Posts và rất nhiều người dùng để lại các nhận xét (comments) về các thông tin được đưa lên, ta thấy đó chính là một kho thông tin vô cùng hưu ích, nếu ta có thể tìm kiếm và phân loại dữ liệu ấy, chúng ta có thể thu được các kết quả khảo sát cần thiết phục vụ cho các hoạt động nghiên cứu hoặc các hoạt động sản xuất kinh doanh. Kết quả khảo sát ấy có thể là những tỉ lệ như "thích" (like) hay là không có ý kiến gì đối với một vấn đề được đưa ra. Việc tìm kiếm và xử lý và tổng hợp các thông tin hưu ích đó cần phải có một mô hình đáp ứng được nhu cầu về việc có thể làm việc trên một lượng dữ liệu lớn và tốc độ cao.

Mô hình MapReduce là một mô hình lập trình giúp các ứng dụng có thể xử lý nhanh hơn một lượng dữ liệu lớn dữ liệu trện các máy phần tán song song, độc lập với nhau từ đó giúp rút ngắn thời gian xử lý toàn bộ dữ liệu. MapReduce có thể chạy trên các phần cứng thông thường (commodity hardware), không đòi hỏi các server chạy MapReduce phải là các máy tính có cấu hình đặc biết mạnh mẽ. Do vậy chi phí triển khai Mapreduce sẽ rẻ hơn.

MapReduce làm đơn giản hóa các giải thuật tính toán phân tán. Với MapReduce, bạn chỉ cần cung cấp hai hàm Map và Reduce cùng với một số thành phần xử lý dữ liệu đầu vào. Do vậy, các nhà phát triển ứng dụng phần tán có thể tập trung nhiều hơn cho phần logic của úng dụng, bỏ qua các chi tiết phức

tạp của việc phân tán xử lý. Sự ra đời của MapReduce đã mở ra cho các doanh nghiện và các trung tâm nhiên cứu cơ hội xử lý các nguồn dữ liệu đồ sộ với chi phí thấp và thời gian nhanh hơn. Hiện nay, đã có nhiều công ty lớn triển khai sử dụng mô hình MapReduce trong việc kinh doanh và khảo sát.

Công ty Amazon sử dụng MapReduce để xử lý các file log trong quá trình mua hàng của khách hàng để dự đoán được xu hướng mua hàng...

Facebook có thể xử lý được khối lượng hơn 10 tỷ hình ảnh mà hộ đang lưu trữ để thu thập các thông tin về hình ảnh, và thu thập 15 terabyte dữ liệu mỗi ngày vào một kho dữ liệu quy mô Petabyte để thực hiện việc khảo sát và đánh giá xu hướng người dùng.

Việc nghiên cứu về xu hướng, đánh giá khảo sát một vấn đề trên quy mô lớn luôn là 1 vấn đề gặp nhiều khó khăn. Trước đây các nhà khảo sát, đánh giá ý kiến trên các đối tượng nghiên cứu thường sử dụng phương pháp thủ công rất tốn kém và mất rất nhiều thời gian để tổng hợp tin tức, chẳng hạn như muốn khảo sát ý kiến của học sinh đối với một số thay đổi trong chương trình học, người ta không thể lựa chọn hỏi ý kiến của tất cả các học sinh mà chỉ có thể lựa chọn một số địa điểm đặc trưng để thực hiện khảo sát, và đôi khi, kết quả của những khảo sát này không mang được tính khách quan vì tâm lý e ngại của các em học sinh. Và những cuộc khảo sát này, đôi khi phải thực hiện trong vòng một vài năm mới có thể có kết quả tổng hợp. Như vậy là mất rất nhiều công sức, của cải và thời gian. Với việc thực trạng hiện nay hầu hết rất cả các em trong lứa tuổi học sinh, sinh viên đều biết sử dụng và thích tham gia các mạng xã hội trên Internet (đặc biết là Facebook) thì việc tìm kiếm một từ khóa có tần suất xuất hiện cao sẽ phản ánh được những xu hướng, những ý kiến của người dùng hơn là việc khảo sát thủ công rất nhiều và việc nhận về những kết quả khảo sát ý kiến. Tổng hợp các thông tin trên máy tính với sự hỗ trợ của mô hình MapReduce sẽ giúp chúng ta có thể thực hiện quá trình đánh giá, khảo sát ý kiến hết sức nhanh chóng và mang lại hiệu quả, cũng như tiết kiệm được rất nhiều thời gian và tiền bạc.

Với những nhu cầu cấp thiết trên, học viên thực hiện nghiên cứu kỹ thuật chỉ mục ngược (Inverted Indexing) đó là phương pháp thực hiện quét một lần trên văn bản sau đó lập danh sách các thuật ngữ (từ, cụm từ) trong file đó và bao gồm cả những thông tin đi kèm với mỗi thuật ngữ (term) (vị trí, tần suất, độ quan trọng...). Các thông tin này sẽ được tổ chức theo một cấu trúc dữ liệu riêng và được gọi là chỉ mục. Với phương pháp đánh chỉ mục ngược kết hợp với mô hình MapReduce sẽ giải quyết được những hạn chế trước đây trong phương pháp thông kê, đánh giá ý kiến trên một quy mô lớn, và đó là lý do học viên lựa chọn