ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Lê Hữu Hảo

ĐÁNH GIÁ VÀ THU THẬP THÔNG TIN TỰ ĐỘNG TRÊN INTERNET SỬ DỤNG DỊCH VỤ TÌM KIẾM

LUẬN VĂN THẠC SỸ

CHUYÊN NGÀNH KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - 2015

ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Lê Hữu Hảo

ĐÁNH GIÁ VÀ THU THẬP THÔNG TIN TỰ ĐỘNG TRÊN INTERNET SỬ DỤNG DỊCH VỤ TÌM KIẾM

Chuyên ngành: Khoa học máy tính

Mã số: **60 48 0101**

LUẬN VĂN THẠC SỸ

CHUYÊN NGÀNH KHOA HỌC MÁY TÍNH

Giáo viên hướng dẫn: TS. Nguyễn Ngọc Hóa

THÁI NGUYÊN - 2015

LỜI CẢM ƠN

Trong thời gian qua, tôi đã nhận được rất nhiều sự hướng dẫn giúp đỡ và động viên tận tình từ nhiều phía. Tất cả những điều đó đã trở thành một nguồn động lực lớn giúp tôi có thể thực hiện được đề tài nghiên cứu được giao. Với tất cả sự cảm kích và trân trọng, tôi xin được gửi lời cảm ơn đến tất cả mọi người.

Trước hết tôi xin chân thành cảm ơn thầy hướng dẫn – Tiến sĩ Nguyễn Ngọc Hóa người đã hết sức nhiệt tình bảo ban hướng dẫn, đóng góp những ý kiến quý báu cho tôi để có thể học tập và hoàn thành luận văn tốt nghiệp này.

Xin gửi lời cảm ơn chân thành nhất đến Ban giám hiệu trường Đại học Công Nghệ Thông Tin và truyền thông – Đại học Thái Nguyên đã tạo điều kiện giúp đỡ tôi có thể thực hiện đề tài. Cảm ơn toàn thể các thầy cô công tác tại trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên đã dạy dỗ và truyền đạt những kiến thức quí báu cho tôi trong suốt thời gian học tập và rèn luyện tại trường.

Tôi xin được gửi lời biết ơn vô hạn tới cha mẹ, người thân đã nuôi dưỡng và tạo điều kiện tốt nhất cho tôi học tập sinh hoạt, ở bên tôi những lúc khó khăn nhất để chuyên tâm thực hiện luận văn.

Cuối cùng, xin cảm ơn tập thể lớp cao học CNTT K12E và đặc biệt những người ban tốt đã ở bên tôi, khuyến khích, động viên tôi và cho tôi những lời khuyên chân thành trong cuộc sống và học tập.

Xin trân trọng cảm ơn!

Thái Nguyên, ngày tháng năm 2015

Học viên

Lê Hữu Hảo

LỜI CAM ĐOAN

Tôi xin cam đoan những nghiên cứu của tôi về "***Đánh giá và thu thập thông tin tự động trên Internet sử dụng dịch vụ tìm kiếm"*** mà tôi viết trong luận văn này là sự thật. Những gì tôi viết ra không sao chép từ các tài liệu, không sử dụng các kết quả của người khác mà không trích dẫn cụ thể.

Tôi xin cam đoan ứng dụng này tôi trình bày trong luận văn là do tôi tự phát triển dưới sự hướng dẫn của thầy Nguyễn Ngọc Hóa, không sao chép mã nguồn của người khác. Nếu sai tôi hoàn toàn chịu trách nhiệm theo quy định của trường Đại học Công nghệ Thông tin và Truyền thông - Đại học Thái Nguyên.

Thái Nguyên, ngày tháng năm 2015

Học viên

Lê Hữu Hảo

MỤC LỤC

DANH MỤC CÁC HÌNH VẼ

DANH MỤC BẢNG BIỂU

GIỚI THIỆU CHUNG

Hiện nay, chúng ta đang sống trong thế kỉ 21, thời đại công nghệ thông tin có những bước phát triển vượt bậc để bắt kịp với xu hướng phát triển chung của xã hội. Kéo theo đó là lượng thông tin khổng lồ về tất cả các lĩnh vực xã hội, chính trị, kinh tế, giải trí, v.v. cũng liên tục phát sinh và phát triển nhanh chóng. Nhu cầu hiểu biết và tìm kiếm của con người ngày càng tăng theo cùng với sự phát triển đó. Tuy nhiên mỗi người, mỗi tổ chức lại có những lĩnh vực quan tâm khác nhau, trình độ hiểu biết về công nghệ thông tin khác nhau, nhu cầu này không chỉ dừng lại ở các cá nhân, mà phổ biến của tất cả mọi người, các tổ chức. Bên cạnh đó lượng thông tin trên Internet cực kì lớn và hỗn tạp từ nhiều nguồn, nhiều loại khác nhau dẫn đến quá trình tìm kiếm khá là khó khăn và vất vả.

Để đáp ứng nhu cầu tìm hiểu của con người, rất nhiều công cụ tìm kiếm đã ra đời cung cấp khả năng tìm kiếm thông tin với tốc độ và phạm vi ngày càng nâng cao và cải thiện. Một trong những công cụ tìm kiếm phổ biến nhất hiện nay đứng đầu là Google, đứng thứ hai là Bing, tiếp đó là Yahoo.... Tuy nhiên các công cụ tìm kiếm trên mang tính chất tức thời, tức là khi người dùng có nhu cầu tìm kiếm thì họ phải trực tiếp vào nhập từ khóa cần tìm và xem kết quả, người dùng sẽ khó nắm bắt và quản lý những thông tin cập nhật thường xuyên liên quan đến những vấn đề nào đó mà mình có nhu cầu theo dõi thường xuyên.

Xuất phát từ thực tế đó, luận văn tốt nghiệp này được hướng đến mục tiêu có thể tự động theo dõi những thông tin mới mà người dùng quan tâm được xuất bản trên nền Web, từ đó xử lý bước đầu để khắc phục tình trạng tránh trùng lặp thông tin và lưu lại trong cơ sở dữ liệu (CSDL) của hệ thống để người dùng có thể dễ dàng tra cứu sau này.

Theo thống kê đã nêu, các công cụ tìm kiếm phổ biến nhất hiện nay có thể đáp ứng được yêu cầu về nguồn dữ liệu lớn, tốc độ phản hồi nhanh và các kết quả có sự liên quan tương đối sát với từ khóa cần tìm kiếm. Vì thế ý tưởng thực hiện của chúng tôi là theo dõi các thông tin trên Internet thông qua việc truy vấn thường xuyên đến các công cụ tìm kiếm phổ biến nhất hiện nay là Google, Bing và Yahoo,

tổng hợp, đánh giác các kết quả sớm nhất và lưu trữ databasse vào cơ sở dữ liệu MongoDB. Như vậy khi người dùng muốn theo dõi các thông tin sẽ truy cập vào hệ thống để có thể xem được những thông tin liên quan mới nhất và nhanh nhất. Để đảm bảo lưu trữ được lượng dữ liệu như vậy, mô hình quản trị cơ sở dữ liệu NoSQL sẽ được nghiên cứu, tìm hiểu để phục vụ việc lưu trữ các tin bài trả về. Phần giao diện tương tác sẽ được thực hiện thông qua ứng dụng Web..

Những kết quả thu được trong luận văn này được tổng hợp trong 3 chương có nội dung sau:

- **Giới thiệu chung**: Trong mục này tôi sẽ giới thiệu chung về luận văn: Nêu thực trạng hiện nay cần giải quyết, từ đó giới thiệu bài toán xuất phát từ nhu cầu thực tế đó, và mục tiêu cũng như những nội dung chính của luận văn.
- **Chương 1**: Tổng quan về tìm kiếm và theo dõi thông tin: Trình bày một số cơ sở lý thuyết chính liên quan đến việc tiến hành các nội dung của luận văn như Search Engine, dữ liệu cấu trúc và cây DOM, thực trạng một số dịch vụ tìm kiếm tiêu biểu hiện nay như Google, Yahoo, Bing,… để lựa chọn triển khai ứng dụng theo dõi thông tin trên Internet.
- **Chương 2**: Mô hình kiến trúc tổng thể và một số thuật toán đánh giá thông tin. Trong chương này sẽ trình bày về mô hình kiến trúc tổng thể của hệ thống cũng như các thành phần chi tiết ở phía máy chủ, máy khách và cơ sở dữ liệu lưu trữ thông tin. Trình bày về các kỹ thuật chính (Framework Strusts, MongoDB, NoSQL...), các phương pháp giải thuật áp dụng cho mỗi phần.
- **Chương 3:** Thực nghiệm ứng dụng đánh giá và thu thập thông tin

  - Trình bày về cách sử dụng hệ thống, những kết quả đã đạt được. Đánh giá các kết quả đã đạt được và những tồn tại của hệ thống.
- **Kết luận chung**

  Tóm tắt lại quá trình xây dựng, những kết quả đã đạt được, ý nghĩa thực tiễn của hệ thống.