# PHIẾU GIAO NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

1. Thông tin về sinh viên
Họ và tên sinh viên: Phạm Nguyễn Tuấn Anh
Điện thoại liên lạc 01234870864 Email: pnta1986@yahoo.com.vn
Lớp: Công nghệ phần mềm – K49 Hệ đào tạo: Chính quy
Đồ án tốt nghiệp được thực hiện tại: trường Đại học Bách Khoa Hà Nội
Thời gian làm ĐATN: Từ ngày 1/3/2009 đến 31/ 5/2009

2. Mục đích nội dung của ĐATN

Mục đích của đồ án: nghiên cứu các hướng tiếp cận trong việc phát hiện các văn bản trùng lặp trong các Search Engine. Áp dụng kết quả nghiên cứu, xây dựng chương trình tìm kiếm các văn bản trùng lặp và đánh giá các kết quả.

3. Các nhiệm vụ cụ thể của ĐATN

- ✓ Tìm hiểu về mô hình của các Search Engine.
- ✓ Nghiên cứu kỹ thuật chỉ mục ngược trong Search Engine.
- ✓ Nghiên cứu các hướng tiếp cận trong việc phát hiện các văn bản trùng lặp.
- ✓ Xây dựng một chương trình cài đặt các kỹ thuật phát hiện văn bản trùng lặp.
- ✓ So sánh và đánh giá từng kỹ thuật.

4. Lời cam đoan của sinh viên:
Tôi - Phạm Nguyễn Tuấn Anh - cam kết ĐATN là công trình nghiên cứu của bản thân tôi dưới sự hướng dẫn của Tiến sỹ Nguyễn Khanh Văn.
Các kết quả nêu trong ĐATN là trung thực, không phải là sao chép toàn văn của bất kỳ công trình nào khác.

*Hà Nội, ngày 24 tháng 5 năm 2009*
Tác giả ĐATN

*Phạm Nguyễn Tuấn Anh*

5. Xác nhận của giáo viên hướng dẫn về mức độ hoàn thành của ĐATN và cho phép bảo vệ:

*Hà Nội, ngày tháng năm*
Giáo viên hướng dẫn

*TS. Nguyễn Khanh Văn*

# TÓM TẮT NỘI DUNG ĐỒ ÁN TỐT NGHIỆP

Các văn bản trùng lặp xuất hiện rất phổ biến, đặc biệt là các văn bản Web. Vấn đề phát hiện các văn bản trùng lặp có ý nghĩa quan trọng đối với nhiều hệ thống. Trong các Search Engine, việc phát hiện được các văn bản trùng nhau sẽ giúp làm tăng hiệu quả lưu trữ và nâng cao chất lượng của các kết quả tìm kiếm. Chính vì thế, phát hiện các văn bản trùng lặp đã trở thành hướng tập trung của các nhà nghiên cứu trong nhiều năm qua. Ở Việt Nam, khi mà các Search Engine bắt đầu xuất hiện trong vài năm gần đây, vấn đề này còn tương đối mới nhưng cũng đã bước đầu được quan tâm.

Đồ án này sẽ tập trung tìm hiểu, so sánh và đánh giá các hướng tiếp cận để giải quyết vấn đề phát hiện các văn bản trùng lặp. Mặt khác, khi áp dụng các phương pháp phát hiện văn bản trùng lặp vào thực tế, ta phải xử lí một lượng văn bản lớn. Vì thế, trong đồ án này, người viết cũng sẽ nêu ra một giải pháp để giải quyết vấn đề này.

Có hai hướng tiếp cận được trình bày trong đồ án: "Phát hiện các văn bản gần trùng" và "Phát hiện sự tái sử dụng các đoạn văn bản cục bộ", trong đó người viết sẽ tập trung vào hướng tiếp cận thứ hai. Hướng tiếp cận này không chỉ cho phép ta xác định độ giống nhau giữa hai văn bản, mà còn chỉ ra được các đoạn văn chung giữa chúng. Các phương pháp và kỹ thuật lấy dấu vân tay trong hướng tiếp cận này sẽ được mô tả, cài đặt và đánh giá để chỉ ra các ưu nhược điểm của chúng. Để giải quyết vấn đề quy mô cuả tập văn bản, kỹ thuật chỉ mục ngược sẽ được áp dụng. Đây là kỹ thuật lưu trữ cơ bản của các Search Enigne, giúp cho việc tìm kiếm nhanh chóng các văn bản trong một tập lớn.

Cuối cùng, với những lý thuyết ở trên, người viết sẽ thử nghiệm xây dựng một chương trình phát hiện các văn bản trùng lặp trên một tập văn bản thực tế. Các kết quả có được từ chương trình sẽ là cơ sở để so sánh và đánh giá các kỹ thuật chống trùng văn bản.

# ABSTRACT OF THESIS

Duplícate documents, especially Web documents, are abundant. Duplicate detection problem is very important for many systems. In Search Engines, successful duplicate detection will increase storing efficiency and enhance quality of search results. Therefore, duplicate document detection has been a focus of researchers for many years. In Vietnam, when Search Engines have just appeared in the recent years, this problem is still new but has begun to be received attention.

The thesis will focus on studying, comparing and assessing approaches to solve duplicate document detection problem. Otherwise, when applying detecting methods in practice, we must handle a huge number of documents. Therefore, in this thesis, the writer will also give a solution to this problem.

There are two approaches which will be represented in this thesis: “Near-Duplicate Detection” and “Local Text Reuse Detection”, and the latter one will be focused by the writer. This approach allows us not only to identify the similarity of two documents, but also to locate the same text between them. Methods and fingerprinting techniques in this approach will be described, setup and assessed to show their own advantages and disadvantages. In order to solve the issue of document collection’s scale, inverted index will be applied. This is the basic storing technique in Search Engines, which allows for rapid searching documents in a big collection.

Finally, with the above theories, the writer will construct a duplicate detecting program for a real document collection. Results from program will be base to compare and assess the fingerprinting techniques.

# LỜI CẢM ƠN

*Trước hết, em xin được chân thành gửi lời cảm ơn sâu sắc tới các thầy cô giáo trong trường Đại học Bách Khoa Hà Nội nói chung và các thầy cô trong khoa Công nghệ Thông tin, bộ môn Công nghệ phần mềm nói riêng đã tận tình giảng dạy, truyền đạt cho em những kiến thức và những kinh nghiệm quý báu trong suốt 5 năm học tập và rèn luyện tại trường Đại học Bách Khoa Hà Nội.*

*Em xin được gửi lời cảm ơn đến TS Nguyễn Khanh Văn - Giảng viên bộ môn Công nghệ phần mềm, khoa Công nghệ Thông tin, trường Đại học Bách Khoa Hà Nội đã hết lòng giúp đỡ, hướng dẫn và chỉ dạy tận tình trong quá trình em làm đồ án tốt nghiệp.*

*Em cũng xin gửi lời cảm ơn đến các anh ở công ty NaiSCorp đã giúp đỡ em rất nhiều về mặt chuyên môn và dữ liệu để em có thể hoàn thành đồ án này.*

*Cuối cùng, em xin được gửi lời cảm ơn chân thành tới gia đình, bạn bè đã quan tâm, động viên, đóng góp ý kiến và giúp đỡ trong quá trình học tập, nghiên cứu và hoàn thành đồ án tốt nghiệp.*

*Hà Nội, ngày 24 tháng 05 năm 2009*

Phạm Nguyễn Tuấn Anh

# MỤC LỤC

# DANH MỤC HÌNH VẼ

# DANH MỤC BẢNG

# DANH MỤC TỪ VIẾT TẮT

| Số TT | Từ | Giải nghĩa |
|---|---|---|
| 1 | NVLV | Người viết luận văn |
| 2 | SE | Search Engine |
| 3 | DVT | Dấu vân tay |
| 4 | NDD | Near-Duplicates Detection |
| 5 | LTRD | Local Text Reuse Detection |
| 6 | DCT | Discrete Cosine Transform |
| 7 | FFT | Fast Fourier Transform |

# ĐẶT VẤN ĐỀ

Hiện tượng văn bản trùng lặp là hiện tượng rất phổ biến trong đời thường. Các văn bản hay đoạn văn bản vì nhiều nguyên nhân thường bị sao chép lại, do đó chúng xuất hiện ở nhiều nguồn khác nhau. Các blogger thường lấy các tin từ các báo điện tử; người gửi thư trả lời thường trích lại một phần hay toàn bộ thư trước; sinh viên viết luận văn sao chép một số phần từ các bài luận văn năm trước… Vì nhiều lí do khác nhau mà người ta muốn tìm và phát hiện được các văn bản trùng nhau. Ví dụ, đối với các hệ thống lưu trữ dữ liệu, các văn bản trùng lặp sẽ làm tốn tài nguyên lưu trữ mà giá trị thông tin mang lại không nhiều. Hay đối với các tổ chức làm việc liên quan đến bản quyền tác giả, họ cần phải tìm ra các văn bản (tác phẩm văn chương, bài báo khoa học…) có sử dụng lại một cách trái phép tác phẩm trước đó. Chính vì vậy, người ta đã quan tâm rất nhiều đến vấn đề phát hiện các văn bản trùng lặp.

Việc phát hiện các văn bản trùng lặp có ý nghĩa rất lớn đối với các Search Engine (SE). Trong quá trình lấy dữ liệu, Crawl Engine (bộ máy lấy dữ liệu từ Internet) sẽ phải bỏ qua các trang Web mà trùng hoặc gần trùng với một trang Web đã được lấy trước đó. Bởi vì các đường link đi ra từ trang Web đó sẽ không khác mấy so với trang Web trước, nên việc phân tích nó là không có ý nghĩa. Điều này sẽ giúp tiết kiệm băng thông, và tránh việc phải gửi request quá nhiều đến host tương ứng. Mặt khác, các SE thường lưu trữ một số lượng khổng lồ các văn bản, dung lượng đến hàng terabyte, mà một phần đáng kể trong số đấy là các văn bản trùng lặp. Điều này làm cho hệ thống trở nên kém hiệu quả.

Tuy nhiên, ảnh hưởng của các văn bản trùng lớn nhất là đến chất lượng của các kết quả tìm kiếm, mối quan tâm hàng đầu của các SE. Người sử dụng khi tìm kiếm bao giờ cũng muốn các kết quả mình mong muốn nhất được đưa ra đầu tiên, và không muốn có nhiều kết quả bị trùng lặp. Nếu như có quá nhiều kết quả trùng, lượng thông tin người dùng nhận được là ít, chất lượng của các kết quả trả về là thấp. Chính vì những lí do trên mà việc phát hiện các văn bản trùng lặp là một trong những vấn đề trọng tâm mà các công ty chuyên về SE nghiên cứu.

Bài toán phát hiện các văn bản trùng lặp là một bài toán khó. Nếu như các văn bản trùng hoàn toàn với nhau (bị sao chép hoàn toàn, không thay đổi) thì chỉ cần một phép kiểm tra đơn giản là ta có thể phát hiện được. Tuy nhiên, phát hiện các văn bản trùng lặp là một vấn đề khó hơn nhiều. Các dạng trùng lặp là vô cùng đa dạng. Một văn bản có thể được sao chép toàn bộ hay chỉ một phần. Các phần văn bản sao chép có thể bị thay đổi (thêm, xoá hoặc bị xáo trộn) và nằm ở những vị trí bất kỳ của văn bản mới. Văn bản mới sau khi sao chép có thể chỉ sai khác với văn bản cũ ở một vài phần nhỏ, ví dụ như 2 trang Web chỉ khác nhau ở đoạn text ghi thời gian hay số người đang truy nhập…, hoặc cũng có thể không giống nhau bao

nhiêu, vì phần chung giữa chúng chỉ là một vài câu hay một đoạn văn nhỏ. Chính vì sự đa dạng trong việc sao chép văn bản mà không thể có một giải thuật hay kỹ thuật nào đo được một cách chính xác sự giống nhau giữa các văn bản.


Vài "chiêu" chống đỡ virus Conflicker - Tin tức
1 Tháng Tư 2009 ... Sâu **Conflicker là gì**? Và nó có thể làm gì? Sâu Conflicker hay còn được gọi là Downadup (hoặc Kido)... những bản vá lỗi từ Microsoft. ...
f.tin247.com/21405767/Vài+"chiêu"+chống+đỡ+virus+Conflicker.html - 60k -
Đã lưu trong bộ nhớ cache - Các trang tương tự

Vài "chiêu" chống đỡ virus Conflicker - HoaBinhToDay.Com - Thành ...
2 bài đăng - Bài đăng mới nhất: 21 Tháng Tư
Sâu **Conflicker là gì**? Và nó có thể làm gì? Sâu Conflicker hay còn được gọi là Downadup (hoặc Kido) cùng với biến thể của.
hoabinhtoday.com/.../14707-vai-oechieua-chong-do-virus-conflicker... - 71k -
Đã lưu trong bộ nhớ cache - Các trang tương tự

Vài "chiêu" chống đỡ virus Conflicker - Diễn đàn Trái Tim Hà Giang ...
23 Tháng Tư 2009 ... Sâu **Conflicker là gì**? Và nó có thể làm gì? Sâu Conflicker hay còn được gọi là Downadup (hoặc Kido) cùng với biến thể của nó tính đến nay đã ...
www.hagiangvui.com/showthread.php?p=594531 - 88k -
Đã lưu trong bộ nhớ cache - Các trang tương tự

Vài "chiêu" chống đỡ virus Conflicker | giatocvip | €¡α †ỘC Vịp ...
Sâu **Conflicker là gì**? Và nó có thể làm gì?Sâu Conflicker hay còn được gọi là Downadup (hoặc Kido) cùng với biến thể của nó tính đến nay đã lây nhiễm một số ...
giatocvip.9kute.com/vai-chieu-chong-do-virus-conflicker.html,... - 56k -
Đã lưu trong bộ nhớ cache - Các trang tương tự

**Hình 1: Một ví dụ cho kết quả tìm kiếm trùng lặp khi tìm kiếm trên Google.**

Khi xây dựng một hệ thống chống trùng văn bản, các SE phải đối mặt với 2 khó khăn chính. Khó khăn đầu tiên và cũng là lớn nhất đó là về quy mô của tập văn bản: các SE lưu trữ một lượng văn bản khổng lồ, lên tới vài chục triệu trang Web. Thứ hai, đó là các SE mỗi ngày tải về hàng nghìn trang Web mới, và chúng phải được chống trùng trước khi được lưu trữ vào trong kho. Do đó, bài toán đặt ra cho các SE là phải làm sao nhanh nhất có thể xác định được một trang Web mới có trùng hoặc gần trùng với một trang Web đã có trong tập hay không. Đây là một vấn đề không dễ giải quyết, cần có những cách xử lí đặc biệt để giải quyết bài toán này.

Các kỹ thuật phát hiện văn bản trùng lặp đã tồn tại là rất phong phú, nhưng có chung một nét cơ bản: chúng đều chia văn bản thành các đoạn nhỏ, rồi dùng một phép biến đổi (thường là dùng hàm băm) để lấy "chữ ký" trên các đoạn văn đó. "Chữ ký" của toàn bộ văn bản được xây dựng từ các "chữ ký" bộ phận. Phép so sánh 2 văn bản sẽ được tiến hành bằng việc đối sánh 2 "chữ ký" văn bản. Mức độ sai khác giữa 2 "chữ ký" sẽ là thước đo cho sự sai khác giữa 2 văn bản. Nói chung, để cho các kết quả so sánh là chính xác, "chữ ký" của văn bản phải biểu diễn được nội dung của văn bản đó nhiều nhất có thể.

Hiện nay có 2 hướng tiếp cận trong việc phát hiện các văn bản trùng lặp: Phát hiện các văn bản gần trùng (*Near-Duplicates Detection* - NND) và Phát hiện sự tái sử dụng đoạn văn cục bộ (*Local Text Reuse Detection* - LTRD). Đây là hai trong số

những phương pháp phát hiện các văn bản trùng lặp mới nhất mà được các nhà nghiên cứu đưa ra. Chúng cũng đã được kiểm nghiệm cho độ chính xác khá cao đối với nhiều loại văn bản.

Ý tưởng chính của phương pháp NND là sử dụng một hàm băm đặc biệt để băm một văn bản thành một giá trị băm tương ứng với văn bản đó. Giá trị băm đó sẽ là "chữ ký" đại diện cho văn bản. Các giá trị băm còn được gọi là dấu vân tay của văn bản. Hàm băm được sử dụng ở đây là **simhash**. Đây là một hàm băm đặc biệt có tính chất là với các văn bản gần trùng các dấu vân tay tương ứng của chúng chỉ sai khác nhau ở số lượng nhỏ các bit. Với tính chất này, việc đo độ giống nhau giữa hai văn bản đơn giản chỉ là xác định số bit khác nhau giữa hai dấu vân tay tương ứng.

Khác với NND, trong LTRD, "chữ ký" của một văn bản là một tập các dấu vân tay. Mỗi dấu vân tay là giá trị băm của một đoạn trong văn bản. Việc so sánh 2 văn bản được thay bằng so sánh 2 tập dấu vân tay tương ứng. Số lượng dấu vân tay chung của cả 2 tập sẽ quyết định độ giống nhau của 2 văn bản. Điểm mạnh của LTRD là ngoài việc xác định mối quan hệ giữa 2 văn bản, phương pháp còn chỉ ra được chính xác các đoạn văn chung giữa 2 văn bản đó.

Để giải quyết vấn đề về quy mô của tập văn bản trong các SE, mỗi phương pháp phát hiện văn bản trùng khác nhau thì có cách làm khác nhau, tuỳ thuộc vào bản chất của các phương pháp. Nhưng nhìn chung, các phương pháp sẽ tìm cách tổ chức việc lưu trữ "chữ ký" của các văn bản một cách tối ưu nhất để sao cho với một "chữ ký" của văn bản mới thì số lượng "chữ ký" của các văn bản có trong tập được chọn để so sánh là ít nhất có thể. Đối với phương pháp NDD, ta xây dựng thành nhiều bảng, với mỗi bảng là tập các dấu vân tay hoán vị của các dấu vân tay đã có ứng với một phép hoán vị. Số lượng các bảng sẽ được tính toán kỹ để sao cho trung hoà được yêu cầu về thời gian tìm kiếm với yêu cầu về dung lượng lưu trữ. Đối với phương pháp LTRD, các tập dấu vân tay của các văn bản được lưu trữ dưới dạng chỉ mục ngược (Inverted Index), là kỹ thuật lưu cơ bản của các SE. Kỹ thuật lưu trữ này tuy đơn giản nhưng mang lại hiệu quả khá cao về mặt thời gian.

Mục đích của luận văn đồ án này trình bày những lý thuyết về 2 hướng tiếp cận trên, với trọng tâm là hướng tiếp cận thứ hai: "*Local Text Reuse Detection*". Các phương pháp và kỹ thuật trong hướng tiếp cận này sẽ được phân tích, so sánh và đánh giá để thấy rõ những ưu và nhược điểm của chúng. Ngoài ra, luận văn đồ án cũng sẽ trình bày kỹ thuật lưu trữ chỉ mục ngược, là kỹ thuật được sử dụng để lưu trữ tập dấu vân tay của các văn bản.

Cuối cùng, người viết sẽ áp dụng các lý thuyết nêu trên để xây dựng một chương trình tìm kiếm các văn bản trùng lặp. Chương trình sử dụng các kỹ thuật phát hiện văn bản trùng lặp khác nhau, và sử dụng kỹ thuật chỉ mục ngược để lưu trữ các dấu vân tay. Chương trình cho phép tìm kiếm các văn bản trùng lặp với một văn bản đầu vào với tập văn bản tìm kiếm lớn và thời gian tìm kiếm ít. Từ những kết quả có

được từ chương trình, người viết sẽ đưa ra những so sánh và đánh giá với từng kỹ thuật để thấy rõ các ưu nhược điểm của chúng.

Với những nội dung trình bày ở trên, luận văn đồ án bao gồm 6 chương, mỗi chương được tóm tắt như sau:

**Chương I**: Giới thiệu tổng quan về mô hình của một SE.

**Chương II**: Trình bày về kỹ thuật lưu trữ theo chỉ mục của các SE.

**Chương III**: Chương này sẽ trình bày hướng tiếp cận: "*Phát hiện các văn bản gần trùng*".

**Chương IV**: Chương này sẽ trình bày hướng tiếp cận: "*Phát hiện sự tái sử dụng đoạn văn bản cục bộ*".

**Chương V**: Trên cơ sở những lý thuyết ở trên, chương này trình bày các bước để xây dựng chương trình tìm kiếm các văn bản trùng lặp.

**Chương VI**: Từ những kết quả có được từ chương trình, người viết sẽ đưa ra những so sánh và đánh giá cho từng kỹ thuật phát hiện văn bản trùng lặp.

**Chương VII**: Tổng kết những công việc đã làm được và nêu lên hướng phát triển của để tài.

Vấn đề phát hiện các văn bản trùng lặp không phải là một vấn đề mới. Trên thế giới, các nhà nghiên cứu đã quan tâm đến vấn đề này từ khá lâu. Trong khoảng 10 năm trở lại đây, số lượng công trình nghiên cứu liên quan đến vấn đề này là rất nhiều. Nguyên nhân chủ yếu chính là do sự phát triển nhanh chóng của các SE. Trong một thập kỷ qua, các SE đã tăng nhanh về số lượng và quy mô của từng SE. Đối với các SE lớn như Google, Yahoo, LiveSearch,… vấn đề chống trùng văn bản luôn được coi là một trong những vấn đề ưu tiên cần nghiên cứu, bởi nó ảnh hưởng trực tiếp tới chất lượng của các kết quả tìm kiếm. Rộng hơn nữa, các kết quả nghiên cứu trong lĩnh vực phát hiện văn bản trùng lặp có thể được áp dụng cho các mục đích khác như: phân cụm văn bản (document clustering), truy xuất dữ liệu có cấu trúc, phát hiện SPAM… Tuy không phải là mới, nhưng ở Việt Nam vẫn chưa có nhiều người quan tâm tới vấn đề phát hiện các văn bản trùng lặp. Sở dĩ là do các động lực để nghiên cứu vấn đề này không nhiều. Vài năm gần đây, khi mà các SE nội bắt đầu xuất hiện thì vấn đề này mới được chú ý hơn, tuy rằng vẫn chỉ ở mức kế thừa các kết quả nghiên cứu trước. Với đề tài đồ án tốt nghiệp "Phát hiện trùng lặp văn bản và ứng dụng vào chỉ mục hiệu quả cho WebCrawler", người viết luận văn mong rằng sẽ đem đến cái nhìn đầu tiên cho người đọc về những hướng tiếp cận nhằm giải quyết một vấn đề còn mới mẻ nhưng cũng rất thú vị này.