Hệ Thống Tìm Kiếm Và So Khớp Tài Liệu Điện Tử Liên Trường Đại Học

Tô Trọng Hiến[1], Nguyễn Việt Phương[1], Nguyễn Hồng Thanh[1], Nguyễn Duy Hoàng[1],

[1] *Trung tâm tính toán hiệu năng cao*
ĐẠI HỌC BÁCH KHOA HÀ NỘI

Tóm tắt **- Hiện nay, nhu cầu tra cứu và chia sẻ tài liệu phục vụ cho công tác nghiên cứu & giảng dạy trong các trường đại học là rất lớn. Bên cạnh đó, các hiện tượng gian lận, sao chép trong học tập (sao chép tài liệu từ khóa trước hoặc thậm chí từ các trường đại học khác) diễn ra ngày một nhiều và ngày càng tinh vi hơn. Điều này ảnh hưởng tiêu cực đến chất lượng tài liệu, luận văn nói riêng, và tới chất lượng giáo dục nói chung. Do vậy, nhu cầu cần có một hệ thống chia sẻ tài liệu vừa hỗ trợ tốt cho công tác nghiên cứu, vừa cung cấp khả năng so khớp, phòng chống gian lận trong học tập là rất cấp thiết.**

Xuất phát từ nhu cầu đó, chúng tôi đề xuất mô hình hệ thống quản lý tài liệu điện tử dựa trên công nghệ lưới dữ liệu [2]. Hệ thống kết nối các tài nguyên tính toán và lưu trữ nằm phân tán của các trường đại học, tạo thành một lưới dữ liệu liên trường đại học. Trên cơ sở đó, hệ thống hỗ trợ người dùng quản lý và chia sẻ, cũng như tìm kiếm và tra cứu tài liệu trong phạm vi trường mình hay trên toàn hệ thống. Đặc biệt, hệ thống còn cung cấp khả năng so khớp tài liệu, dựa trên giải thuật PLSA [6]– một hướng tiếp cận theo ngữ nghĩa. Với mỗi tài liệu đầu vào, hệ thống sẽ dựa trên kho tài liệu hiện có để phân tích sự trùng khớp về mặt nội dung, và kết luận bao nhiêu phần trăm tài liệu là sao chép từ các tài liệu khác. Bên cạnh đó, chúng tôi còn tiến hành những thử nghiệm phân nhóm tài liệu theo lĩnh vực. Kết quả cho thấy hệ thống đã nâng cao được khả năng đáp ứng cũng như tính chính xác của việc tìm kiếm và so khớp tài liệu. Hệ thống được xây dựng đảm bảo các tính chất: an toàn, sẵn sàng, và khả năng mở rộng cao. Hệ thống hiện đang được triển khai thử nghiệm tại Trung tâm Tính Toán Hiệu Năng Cao, Đại Học Bách Khoa Hà nội.

Từ khóa **– Tính toán lưới, so khớp tài liệu, tìm kiếm phân tán, bảo mật lưới**

I. GIỚI THIỆU CHUNG

Vấn đề quản lý tài liệu điện tử nằm phân tán trên các trường đại học đặt ra rất nhiều thách thức. Đặc điểm của các tài liệu điện tử đó là các tài nguyên động. Tính chất động không chỉ thể hiện về mặt dữ liệu (các tài liệu có thể được thêm, bớt, hay sửa đổi), mà còn về mặt sở hữu. Các tài liệu thuộc sở hữu của các cá nhân nằm trong các tổ chức khác nhau. Mỗi tổ chức lại có các chính sách quản lý tài nguyên và truy cập riêng. Trong khi đó, yêu cầu về một hệ thống quản lý tài liệu phân tán trên mạng đòi hỏi tính mềm dẻo (khả năng dễ dàng thêm bớt các tài nguyên lưu trữ, tính toán), tính bảo mật cho dữ liệu thuộc các tổ chức, và tính cộng tác chia sẻ dữ liệu trong các nhóm hay trong các lĩnh vực nghiên cứu. Việc giải quyết các yêu cầu đặt ra này trên một tập các tài nguyên động là rất phức tạp.

Trong những năm trở lại đây, tính toán lưới [1] nói chung và lưới dữ liệu [2] nói riêng đã có những bước tiến không ngừng trên phạm vi toàn thế giới. Công nghệ tính toán lưới cho phép kết hợp sức mạnh tính toán của nhiều máy tính đơn lẻ, tạo thành sức mạnh tính toán tổng hợp. Do vậy, tính toán lưới là một hướng đi đầy triển vọng để giải quyết các vấn đề về tính toán phân tán nói chung, cũng như bài toán quản lý tài liệu điện tử trên mạng nói riêng.

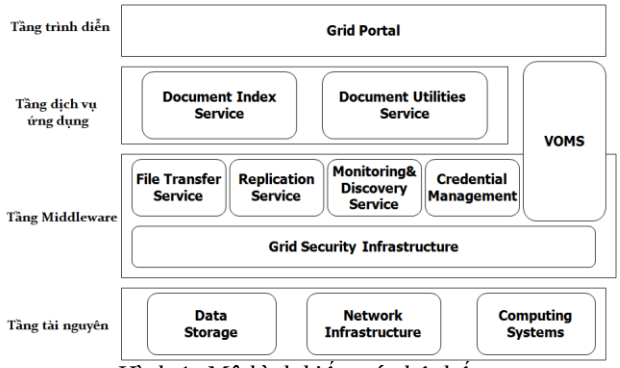
Trong bài báo này, chúng tôi trình bày Giải pháp quản lý tài liệu điện tử nằm phân tán trong các trường đại học, đồng thời đưa vào hệ thống các phương pháp tìm kiếm, so khớp văn bản theo hướng tiếp cận ngữ nghĩa nhằm nâng cao hiệu quả việc tra cứu tài liệu và hỗ trợ giảng viên trong việc phát hiện các gian lận, sao chép trong học tập. Trong các phần tiếp theo của bài báo, chúng tôi xin lần lượt trình bày về mô hình kiến trúc của hệ thống, tiếp đó là vấn đề về quản lý đăng ký lưới, dịch vụ lưới dữ liệu, bài toán tìm kiếm trong môi trường lưới, và bài toán so khớp tài liệu chúng tôi đã thực hiện.

II. MÔ HÌNH KIẾN TRÚC HỆ THỐNG

Giải pháp đưa ra kết nối các tài nguyên tính toán, lưu trữ của các trường đại học thành một lưới dữ liệu, mỗi trường là một nút lưới. Người dùng truy cập và sử dụng các chức năng của hệ thống thông qua cổng thông tin của mỗi trường. Hình 1 minh họa Mô hình kiến trúc của hệ thống

Kiến trúc hệ thống gồm 4 tầng:

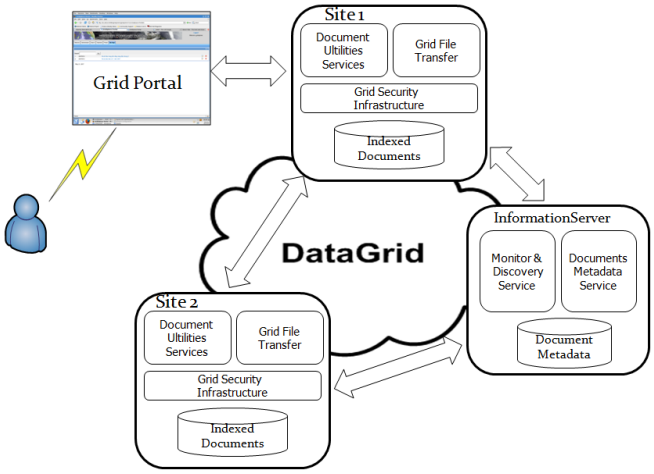
- Tầng tài nguyên: là tầng thấp nhất của hệ thống, bao gồm tất cả các tài nguyên cơ bản, như các hệ thống tính toán, hệ thống lưu trữ của các trường đại học cũng như hạ tầng mạng kết nối các tài nguyên lại với nhau về mặt vật lý.
- Tầng Middleware (hay tầng phần mềm nền): bao gồm rất nhiều các dịch vụ, các giao diện lập trình ứng dụng (APIs), cung cấp khả năng quản lý các tài nguyên tính toán & lưu trữ nằm phân tán. Tầng middleware giúp triển khai hạ tầng lưới một cách toàn diện, tạo cơ sở để phát triển những ứng dụng trên nền lưới. Hiện nay, Globus Toolkit là một trong những grid middleware được sử dụng rộng rãi nhất trong các dự án tính toán lưới.



Hình 1- Mô hình kiến trúc hệ thống

- Tầng dịch vụ ứng dụng: tầng dịch vụ ứng dụng cung cấp các dịch vụ lưới hướng ứng dụng & hướng người dùng, nhằm tận dụng sức mạnh mà hạ tầng lưới mang lại. Trong phạm vi hệ thống quản lý tài liệu điện tử, tầng ứng dụng triển khai các dịch vụ tiện ích như quản lý tài liệu, tìm kiếm, đánh chỉ mục, so khớp văn bản…
- Tầng trình diễn: được triển khai theo mô hình cổng thông tin trên nền lưới (grid portal) [4], kết nối người dùng với hệ thống lưới. Cung cấp một điểm truy nhập chung nhất, thông qua giao diện web, cho toàn bộ các tài nguyên tính toán, lưu trữ phân tán, cũng như các dịch vụ lưới tiện ích của hệ thống. Có thể nói, tầng trình diễn làm trong suốt sự phức tạp của lưới tới người dùng.

Hình 2 minh họa Mô hình triển khai của hệ thống:

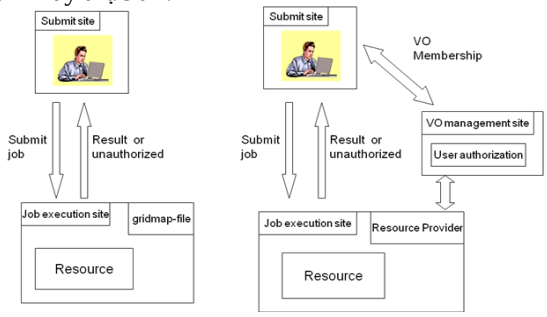


Hình 2 – Mô hình triển khai hệ thống

Người dùng truy cập hệ thống thông qua portal của các trường, từ đó kết nối đến các dịch vụ lưới trên các site cục bộ và thực hiện các yêu cầu như tìm kiếm, so khớp văn bản…Với mỗi yêu cầu của người dùng, site cục bộ sẽ liên hệ với Information Server, thông qua Dịch vụ Giám sát & quản lý thông tin toàn hệ thống (Monitoring & Discovery Service) để lấy về thông tin các site khác trên hệ thống & đồng thời gửi các yêu cầu xử lý (như tìm kiếm hay so khớp một văn bản) tới các site đó. Sau khi nhận được kết quả từ các site khác, site cục bộ phải tổng hợp kết quả và trả về cho người dùng qua portal. Việc tổng hợp kết quả cần phải có các thông tin về dữ liệu mô tả từ các site, được cung cấp bởi Dịch vụ siêu dữ liệu mô tả từ Information Server.

III. MÔ HÌNH QUẢN LÝ ĐĂNG KÝ LƯỚI.

Khi một site mới gia nhập lưới, ngoài việc phải triển khai các hạ tầng tính toán lưới như Global Toolkit, Portal ..., site đó còn phải đăng ký hoạt động với lưới, thông qua việc đăng ký người dùng lưới mới, được thực hiện bởi các người quản trị của mỗi site. Ở các hệ thống lưới vừa và nhỏ, người dùng lưới sẽ phải tự đăng ký sử dụng tài nguyên với từng cụm máy thực thi. Quá trình này là thủ công, đòi hỏi tốn nhiều thời gian và công sức cho cả phía người dùng cũng như người quản trị các cụm máy thực thi.



Hình 3 – Mô hình quản lý người dùng với VOs

Ở hệ thống quản lý tài liệu, nơi có sự tham gia của nhiều đơn vị, kéo theo số lượng lớn người dùng cũng như nhiều loại tài nguyên, thì quy trình thủ công như vậy không thể đáp ứng được. Yêu cầu đặt ra là: cần phải nâng mức quản lý người dùng lên 1 cấp (Hình 3) cách gom nhóm việc đăng ký sử dụng tài nguyên vào một đầu mối thống nhất cho toàn bộ lưới, hệ thống sẽ phải tự động cập nhật các người dùng trên site này lên toàn bộ hệ thống tài nguyên bên dưới. remove đi – tớ thấy ko cần quy trình đăng ký và cập nhật phải đảm bảo tính trong suốt cho lưới với người dùng. Từ các yêu cầu trên, giải pháp được lựa chọn là: hệ thống quản lý thành viên tổ chức ảo VOMS (Virtual Organization Membership Service) [3].

Xét trên quan điểm của VO (Virtual Organization), việc người dùng tham gia lưới cũng chính là tham gia VO của tổ chức lưới; việc người dùng đăng ký lưới cũng chính là đăng ký vào VO. Để đáp ứng các yêu cầu về tính trong suốt với lưới cho người dùng, VOMS, hệ thống quản lý thành viên tổ chức ảo, được phát triển để hiện thực hóa quy trình quản lý người dùng trong môi trường lưới.

Quy trình đăng ký vào lưới bao gồm 5 bước như sau:

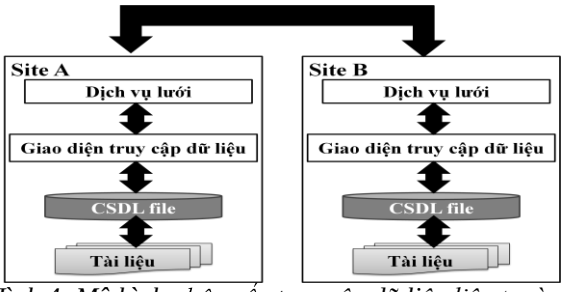
1. Bước chuẩn bị: Người dùng sở hữu cặp khóa public-private key được các CA mà VOMS server tin tưởng chứng nhận. (các CA hiện tại được chưa giới thiệu goodas là gì? tin tưởng: tớ thấy ko cần phải ví dụ thế này: vì ví dụ phải làm cho người dùng dễ hiểu hơn. đằng này Hanoi CA là gì? AGP CA là gì?). Sau đó, người dùng convert cặp khóa public-private key sang định dạng pkcs12 và import vào trình duyệt.
2. Người dùng truy cập vào VOMS Server của hệ thống, hệ thống yêu cầu người dùng tuân thủ các quy định sử dụng lưới của VO, cung cấp các thông tin email, và một số thông tin cá nhân khác ...
3. VOMS Server gửi email xác nhận đến cho người dùng, người dùng xác nhận có muốn tham gia VO hay từ chối.
4. Nếu người dùng xác nhận, VO Admin nhận được thông báo có người dùng muốn gia nhập VO và sẽ quyết định có cho phép người dùng gia nhập lưới
5. Nếu người dùng được VO Admin chấp nhận, họ chính thức là **thành viên của VO**. Trên các máy thực thi, định kỳ sẽ cập nhật các thành viên trong VO và ánh xạ họ vào các account cục bộ trên máy thực thi. Người dùng chính

thức có thể sử dụng các tài nguyên trên máy thực thi hay còn gọi là **người dùng lưới.**

Với việc áp dụng VOMS, quy trình đăng ký mới một site với lưới được đơn giản hóa rất nhiều. Người quản trị của site đó chỉ việc truy cập vào trang đăng ký của VOMS Server của hệ thống để đăng ký người dùng lưới cho site mình, hệ thống sẽ tự động cập nhật thông tin người dùng lưới mới trên tất cả các trạm thực thi khác trên lưới.

IV. LƯỚI DỮ LIỆU HỆ THỐNG

Dữ liệu văn bản trong hệ thống được lưu trữ phân tán trên nhiều trường, do đó vấn đề là phải xây dựng một tầng dịch vụ lưu trữ dựa trên nền tảng lưới dữ liệu [2] remove đi, thừa ☺. Dịch vụ lưu trữ này kết nối các tài nguyên từ các trường, tạo thành một kho tài liệu chung nhất và trong suốt với người sử dụng, qua đó cung cấp các tiện ích trên dữ liệu như tìm kiếm, so khớp, download, upload... Các dịch vụ trên mỗi site kết nối với nhau qua môi trường lưới cho phép người dùng tiếp cận thông tin trên toàn hệ thống mà chỉ thông qua một giao diện dịch vụ duy nhất.
xóa đi. Triển khai không đúng



Hình 4: Mô hình phân cấp truy cập dữ liệu liên trường

Tầng dịch vụ lưới: cài đặt các dịch vụ lưới như xóa điphát hiện tài nguyên, dịch vụ tìm kiếm và so khớp. Các dịch vụ này sử dụng các giao diện truy cập CSDL

Tầng giao diện truy cập dữ liệu: cung cấp các phương thức truy cập cơ sở dữ liệu: truy xuất, cập nhật …

CSDL file: thông tin về nên để : người dùng, tài liệu và nhóm tài liệu

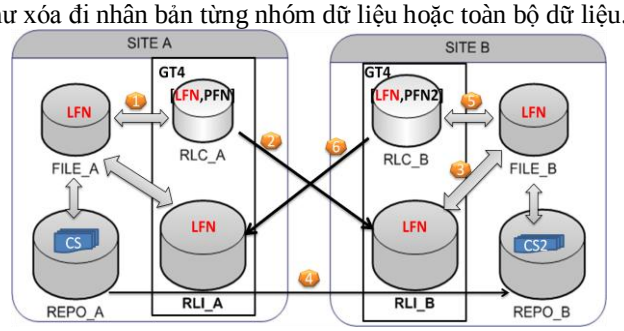
Kho tài liệu: lưu trữ vật lý tài liệu theo nhóm thư mục

Đảo lại thứ tự trình bày: Kho dữ liệu → dịch vụ lưới (để phù hợp với mô hình tổng quan ở trên)
chỗ này in nghiêng

Trong thực tế một hệ thống phân tán chia sẻ tài liệu gặp nhiều rủi do liên quan đến dữ liệu như máy lưu trữ bị lỗi. Việc này làm cho các chức năng của hệ thống không thực hiện được hoặc thực hiện thiếu chính xác. Để giải quyết vấn đề này chúng tôi đề xuất mô hình nhân bản dữ liệu (data replication) trên môi trường lưới:

Dịch vụ định vị bản sao RLS (Replica Location Service) lưu giữ thông tin về các bản sao. Mỗi khoản mục trong kho tương ứng với một tên tệp lôgic - LFN. Tương ứng với mỗi tệp hoặc một nhóm tệp lôgic là một hoặc một nhóm địa chỉ vật lý - PFN. Dịch vụ quản lý ánh xạ giữa tên lôgic với địa chỉ vật lý lưu trữ tệp đó. Dịch vụ RLS tại mỗi site có hai chế độ là quản lý địa phương RLC (Local Replica Database) và quản lý toàn cục RLI (Replica Index Service). Mỗi RLC có thể đăng ký một RLI, ngược lại RLI có thể đăng ký nhiều RLC. Mô hình này giúp tạo bản sao tại nhiều trường.

Để đảm bảo tính tự trị của trường thì việc nhân bản là do các trường tự nguyện. Sau khi quản trị của hai trường thỏa thuận với nhau, họ có thể tạo ☺ bản sao tại nhiều trường cũng như xóa đi nhân bản từng nhóm dữ liệu hoặc toàn bộ dữ liệu.



Hình 5: mô hình nhân bản dữ liệu

Site A xóa đi tạo một bản sao của nhóm tin học sang site B:

1. Site A liệt kê các cặp (LFN, PFN) từ CSDL FILE_A của các file thuộc nhóm tin học, sau đó tạo ánh xạ bản sao cụ bộ tại RLC_A: [LFN, PFN]
2. Sau khi bên A cấu hình để RLC_A trỏ vào RLI_B, dữ liệu LFN của RLC_A được tự động cập nhật sang RLI_B
3. Có một dịch vụ bên site B được kích hoạt khi có dữ liệu cập nhật vào xóa RLI_B: tìm các LFN mới được cập nhật bằng cách đối chiếu các LFN của RLI_B với các LFN trích suất từ CSDL FILE_B.
4. Sau đó site B sử dụng giao thức GridFTP copy các file có LFN mới từ bên site A vào thư mục với chủ sở hữu là quản trị bên B, và quyền truy nhập là public
5. Site B tạo ánh xạ trong CSDL bản sao cục bộ RLC_B: [LFN, PFN2]
6. Sau khi được cấu hình để RLC_B trỏ vào RLI_A, dữ liệu LFN của RLC_B được tự động cập nhật sang RLI_A

Nhờ khả năng nhân bản dữ liệu mà hệ thống không những tăng độ tin cậy mà còn tăng hiệu năng trong trường hợp download dữ liệu.

Trong phần này, chúng tôi đã trình bày về mô hình dịch vụ lưới dữ liệu của hệ thống, trên đó triển khai các dịch vụ chính của hệ thống, như tìm kiếm, so khớp, quản lý, download, upload tài liệu... Trong các phần tiếp theo, chúng tôi xin đi sâu vào các giải thuật tìm kiếm và so khớp được triển khai trên môi trường lưới dữ liệuxóa dấucũng như những thử nghiệm của chúng tôi trong việc phân loại tài liệu nhằm đạt hiệu quả và độ chính xác cao hơn trong quá trình tìm kiếm & so khớp.

V. BÀI TOÁN TÌM KIẾM

Hệ thống của chúng tôi sử dụng một phương pháp trích rút thông tin khá cổ điển và hiện vẫn được sử dụng rộng rãi, đó là mô hình không gian vector.

Trong mô hình không gian vector (Vector Space Model – VSM), văn bản và các câu truy vấn được biểu diễn bằng các vector từ khóa trong không gian nhiều chiều. Với phương pháp biểu diễn này, việc tìm các văn bản có liên quan tới truy vấn được thực hiện bằng cách tính độ tương tự giữa các vector tương ứng, một phương pháp được sử dụng là tính cosin của góc giữa vector văn bản và vector truy vấn:

$$\cos\theta = \frac{d_i.q}{|d_i|.|q|} = \frac{\sum_j w_{i,j} * w_{q,j}}{\sqrt{\sum_j w_{i,j}^2}\sqrt{\sum_j w_{q,j}^2}}$$

Trong đó di là văn bản và q là câu truy vấn; $w_{i,j}$ và $w_{q,j}$ lần lượt là trọng số của từ khóa qj đối với văn bản di và câu truy vấn q. Một phương pháp tính trọng số của từ khóa được sử dụng rộng rãi trong mô hình không gian vector đó là phương pháp TF*IDF. Trong phương pháp này, trọng số của một từ khóa được tính bởi tích giữa tần suất xuất hiện của từ khóa đó trong một văn bản (TF - Term Frequency) với nghịch đảo của số văn bản chứa từ khóa đó trong tập các văn bản đã có (IDF – Inverse Document Frequency).

Tổng hợp kết quả tìm kiếm.

Một vấn đề đặt ra cho bài toán tìm kiếm trong môi trường lưới đó là: các tài liệu được lưu trữ, đánh chỉ mục cũng như tìm kiếm được thực hiện phân tán trên các site. Do vậy, cần phải có một phương pháp để tổng hợp kết quả tìm kiếm từ các site và đưa ra cho người dùng. Phương pháp này phải đảm bảo kết quả tìm kiếm tổng hợp được tớ thấy tốt = hơi chung chung. viết thằng là tốt hơn tập trung và dữ liệu cần thiết cho quá trình tổng hợp phải không lớn để đảm bảo cho hiệu năng hoạt động & chất lượng đường truyền của hệ thống.

Chúng tôi lựa chọn phương pháp CVV (Cue-Validity Method) [7] tổng hợp kết quả tìm kiếm từ các site cục bộ, dựa trên việc xếp hạng các site tìm kiếm. Trong phương pháp này, ứng với mỗi câu truy vấn đầu vào, mỗi site sẽ được tính toán một giá trị trọng số riêng (hay còn gọi là Goodness Score – $G_{i,q}$), các trọng số này được áp dụng trong quá trình tổng hợp kết quả tìm kiếm, mà thực chất là việc chuyển đổi từ các kết quả tìm kiếm cục bộ sang toàn cục để có thể trộn lại, sắp xếp và trả về cho người dùng. Ưu điểm của phương pháp CVV khi áp dụng đối với bài toán tìm kiếm phân tán trên môi trường lưới đó là: dữ liệu duy nhất cần thiết cho quá trình tính toán trọng số cho các site tìm kiếm đó là DF (Document Frequency) từ mỗi site. Thành phần dữ liệu này là không lớn và có thể dễ dàng kết xuất trong quá trình đánh chỉ mục tài liệu trên mỗi site theo phương pháp VSM.

Giá trị Goodness Score của mỗi site ứng với một câu truy vấn đầu vào q được tính như sau:

$$G_{i,q} = \sum_{j=1}^{M} CVV_j * DF_{i,j}$$

Trong đó M là số lượng từ khóa trong câu truy vấn, $DF_{i,j}$ là giá trị DF của từ khóa thứ j trong tập hợp văn bản của site thứ i. CVV_j là phương sai của giá trị trọng số CVj tương ứng với từ khóa thứ j trên tất cả các site. Giá trị CVj được tính từ $CVV_{i,j}$ của từ khóa đó đối với tập các văn bản trên site i – c_i. CVV_j đánh giá mức độ của phân biệt của từ khóa j đối với các văn bản trong tập c_i so với các tập văn bản khác, và được tính như sau:

$$CV_{i,j} = \frac{\frac{DF_{i,j}}{N_i}}{\frac{DF_{i,j}}{N_i} + \frac{\sum_{k\neq i}^{|C|} DF_{k,j}}{\sum_{k\neq i}^{|C|} N_k}}$$

Trong đó, N_i là số văn bản trong tập hợp c_i của mỗi site, và |C| là tổng số site của hệ thống. Phương sai CVV_j của $CV_{i,j}$ đánh giá mức độ chênh lệch của sự phân tán xuất hiện của từ khóa j trên các tập văn bản và có thể được sử dụng để đánh giá mức độ hữu dụng của từ khóa j trong việc phân biệt một tập hợp văn bản này với tập hợp khác. Giá trị CVV_j càng lớn thì từ khóa đó càng hữu dụng:

$$CVV_j = \frac{\sum_{i=1}^{|C|}(CV_{i,j} - \overline{CV_j})^2}{|C|}$$

Trong đó $\overline{CV_j}$ là giá trị trung bình của các $CV_{i,j}$ trên tất cả các tập văn bản: $\overline{CV_j} = \frac{\sum_{i=1}^{|C|} CV_{i,j}}{|C|}$

Có thể thấy các giá trị $G_{i,q}$ thu được không cho ta biết chính xác có bao nhiêu văn bản liên quan đến câu truy vấn q trong tập ci hay mức độ liên quan của các văn bản đó với câu truy vấn, $G_{i,q}$ chỉ cho ta biết trong số |C| các tập văn bản, các từ khóa trong câu truy vấn phân bố tập trung nhất ở tập hợp nào.

Các kết quả thử nghiệm cho thấy, phương pháp tổng hợp kết quả này có thời gian xử lý tổng hợp nhanh, và đặc biệt lượng dữ liệu yêu cầu trong quá trình tổng hợp là không nhiều, từ đó giúp tối ưu được băng thông hệ thống trong quá trình tổng hợp kết quả. Dịch vụ cung cấp thông tin (tính toán các giá trị Goodness) phục vụ quá trình tổng hợp được triển khai trên Information Server (hay còn gọi là Broker Server) của hệ thống.

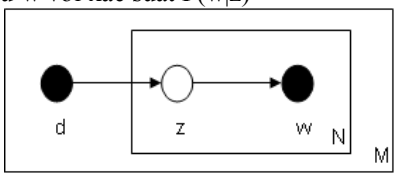
VI. BÀI TOÁN SO KHỚP TÀI LIỆU

Trong vấn đề so khớp tài liệu dựa trên nội dung, chúng tôi sử dụng phương pháp PLSA (Probabilistic Latent Semantic Analyis) [6] để trích rút được những nội dung chính của các tài liệu trước khi so khớp. Ý tưởng chính của phương pháp này là giả thuyết mỗi một tài liệu là sự tổ hợp của một tập các chủ đề (topic) ẩn z với các trọng số p (z | d) (là xác suất xuất hiện của chủ đề z trong tài liệu d), trong đó mỗi chủ đề lại là sự tổ hợp của một tập các từ vựng xuất hiện trong các tài liệu với các trọng số p(w|z) (là xác suất xuất hiện của từ vựng w trong chủ đề z). Khi đó, dựa trên các thông tin về các chủ đề của từng tài liệu, cụ thể ở đây là p(w|z), chúng ta hoàn toàn có thể so khớp nội dung của tài liệu truy vấn với các tài liệu nằm trong cơ sở dữ liệu.

Mô hình tổng quát

Ý tưởng của PLSA dựa trên mô hình xác suất Aspect hay còn gọi là mô hình biến ẩn cho tập dữ liệu kép (w, d), trong đó mỗi xuất hiện (w, d) đều tương tác với một tập biến ẩn z. Mô hình Aspect được mô tả như sau:

- Chọn một tài liệu d với xác suất P(d)
- Chọn một chủ đề ẩn z với xác suất P(z|d)
- Sinh một từ w với xác suất P(w|z)



Thuật toán

Mục tiêu của PLSA là tìm các tham số P(w|z) và P(z|d) cho mô hình Aspect của tập các document. Sử dụng phương pháp EM (Expectation Maximization) ta có thuật toán:

Khởi tạo P(w|z) và P(z|d)

Vòng lặp

Tính $P(z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{l=1}^{K} P(w_j|z_l)P(z_l|d_i)}$

Tính $P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i,w_j)P(z_k|d_i,w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i,w_m)P(z_k|d_i,w_m)}$

Tính $P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i,w_j)P(z_k|d_i,w_j)}{n(d_i)}$

Kiểm tra điều kiện dừng của vòng lặp

Việc so khớp tài liệu dựa trên những thông tin P(z|d) mà ta có được khi phân tích các tài liệu. Với hai tài liệu d và q bất kì, mức độ giống nhau giữa chúng được tính bởi công thức cos giữa 2 vector:

$$s(d,q) = \frac{\sum_i d_i q_i}{\sqrt{\sum_i d_i^2} \cdot \sqrt{\sum_i q_i^2}}$$

Với $d_i = P(z_k|d)$ và $q_i = P(z_k|q)$

Nhận xét

Phương pháp PLSA giải quyết được vấn đề từ đa nghĩa và từ đồng nghĩa khi phân tích nội dung của các tài liệu.

Hiệu quả của phương pháp phụ thuộc vào số chủ đề được thiết lập, số chủ đề càng lớn thì độ chính xác càng cao.Vì thế, nếu cơ sở dữ liệu được chia nhỏ và chia thành các nhóm có cùng nội dung thì PLSA hoạt động càng tốt.Như vậy, cơ sở dữ liệu được phân tán trên môi trường lưới và lưu trữ thành các nhóm sẽ giúp cho việc so khớp và tìm kiếm tài liệu dựa trên nội dung bằng phương pháp PLSA đưa ra kết quả tốt hơn so với việc cơ sở dữ liệu được lưu trữ tập trung.

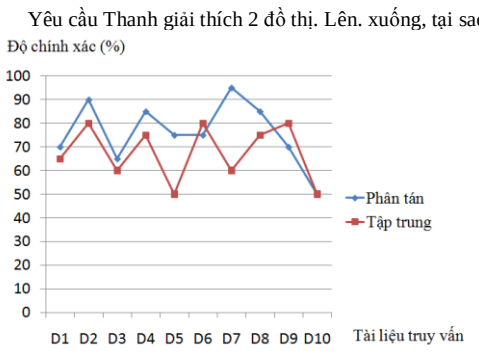
Kết quả thực nghiệm trên hệ thống

Chúng tôi thực nghiệm với tập cơ sở dữ liệu gồm 240 tài liệu được lấy từ abstract của 240 bài báo của trang web http://ieee.org/. Thuật toán được thực nghiệm trên hệ thống theo hai bước: khi bộ dữ liệu được lưu trữ tập trung trên một site và khi bộ dữ liệu được lưu trữ phân tán trên nhiều site khác nhau và được phân loại theo các lĩnh vực khác nhau.

Để tính toán độ chính xác của thuật toán, chúng tôi chọn 10 tài liệu dùng để truy vấn.Với mỗi tài liệu này, chọn 20 tài liệu có nội dung liên quan đến nó nhất để tạo bộ kết quả chuẩn. Sử dụng phương pháp PLSA để so khớp nội dung của tài liệu truy vấn với bộ CSDL và lấy ra 20 tài liệu cho kết quả so sánh có tỉ lệ nội dung giống với tài liệu truy vấn nhất. Sau đó, so sánh kết quả thu được với bộ kết quả chuẩn để tính độ chính xác của PLSA theo công thức

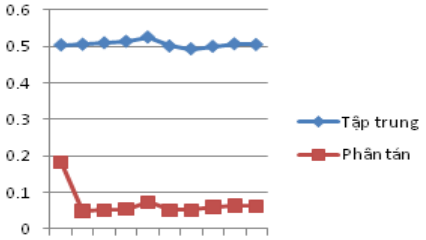
$$\text{AccuracyRate} = \frac{\text{Số tài liệu giống với bộ kết quả chuẩn}}{\text{20 (là tổng số tài liệu có trong kết quả chuẩn)}}$$

Dựa vào kết quả thực nghiệm cho thấy trong hầu hết trường hợp, việc xử lý dữ liệu phân tán double space --& phân nhóm đều cho kết quả về độ chính xác tốt hơn so với khi xử lý dữ liệu tập trung:



Biểu đồ 1 – Kết quả thử nghiệm về độ chính xác

Tại sao lại có sự chênh lệch thế này? Chắc chắn nó có liên quan đến số Group/trạm mà mình phân ra. Tớ thấy cái này quan trọng, vì mấy ai quan tâm đến mấy cái công thức



phức tạp kia.

kết quả tốc độ mình vẫn lấy kết quả test trước đây. VẪN THIẾU ĐƠN VỊ 2 TRỤC NÈ ☺

Ngoài ra khi dữ liệu được lưu phân tán, quá trình tìm kiếm và so khớp dữ liệu online cho tốc độ nhanh hơn so với lưu trữ dữ liệu tập trung do mỗi nút lưới chỉ xử lý phần dữ liệu tại nút đó trong khi nếu lưu tập trung hệ thống phải xử lý một khối lượng tài liệu lớn hơn rất nhiều.

VII. KẾT LUẬN

Trong bài báo này, chúng tôi đã đề xuất giải pháp lưới dữ liệu, xây dựng hệ thống tìm kiếm và so khớp tài liệu điện tử liên trường đại học. Đồng thời đưa ra những thử nghiệm về phân loại tài liệu nhằm đạt hiệu quả và độ chính xác cao hơn trong quá trình tìm kiếm & so khớp. Những kết quả thử nghiệm và chạy thử hệ thống cho thấy: hệ thống không chỉ đã đáp ứng được các yêu cầu đặt ra về tính mềm dẻo, tính bảo mật, và khả năng cộng tác, chia sẻ tài nguyên giữa các trường đại học; mà còn hỗ trợ một cách tích cực cho việc tra cứu tài liệu nghiên cứu cũng như so khớp tài liệu và phòng chống gian lận trong học tập.

VIII. TÀI LIỆU THAM KHẢO

[1]. *Viktors Berstis, "Fundamentals of Grid Computing", IBM Redbooks, 2002.*

[2]. *Ann Chervenak, Ian Foster, Carl Kesselman, Charles Salisbury, and Steven Tuecke, "The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets", Journal of Network and Computer Applications, 2001.*

[3]. *R. Alfieri, R. Cecchini, V. Ciaschini, L. dell'Agnello, A´. Frohner, A. Gianoli, K. L¨orentey, and F. Spataro, "VOMS, an Authorization System for Virtual Organizations" , 2003.*

[4]. *JasonNovotny,MichaelRussell,OliverWehrens, " GridSphere: An Advanced Portal Framework"*

[5]. *Borja Sotomayor, Lisa Childers, "Globus®Toolkit 4 Programming Java Services", 2005*

[6]. *Thomas Hofmann, "Probabilistic Latent Semantic Analysis", EECS Department, University of California, Berkeley, 1999*

[7]. *Budi Yuwono, Dik L. Lee,"Server ranking For Distributed Text Retrieval Systems on the Internet",1997*