Giải Pháp Lưới Dữ Liệu Quản Lý Tài Liệu Điện Tử Được Lưu Trữ Phân Tán Trong Tổ Chức ảo

Lê Đức Tùng[1], Nguyễn Thanh Thủy[1,2]
Đào Quang Minh[1]
Tô Trọng Hiến[1,] Nguyễn Việt Phương[1], Nguyễn Hồng Thanh[1], Nguyễn Duy Hoàng[1],

[1] *Trung tâm tính toán hiệu năng cao*

[2] *Bộ môn hệ thống thông tin, Viện Công nghệ thông tin truyền thông*

ĐẠI HỌC BÁCH KHOA HÀ NỘI

***Tóm tắt* - Tính toán lưới [1] hiện đang được sử dụng rộng rãi trong các ứng dụng quản lý có tính chất phân tán về tài nguyên. Các công nghệ trong môi trường lưới được chuẩn hóa tạo nên cơ sở hạ tầng lưới bảo mật và tin cậy. Trong bài báo này chúng tôi đề xuất ra mô hình quản lý tài liệu dựa trên công nghệ lưới. Mô hình phải đảm bảo các tính chất sau: 1) An toàn: dữ liệu trong mô hình chỉ được truy cập bởi người dùng có thẩm quyền, và phân cấp mô hình truy cập; 2) Sẵn có: Dữ liệu luôn sẵn sàng với các truy vấn đến hệ thống, các kịch bản sao lưu đưa ra nhằm mục đích tạo nên tính sẵn sàng; 3) Co dãn: Khả năng dễ dàng thêm bớt các tài nguyên lưu trữ mà không làm ảnh hưởng đến hoạt động của hệ thống, tài nguyên lưu trữ có thể là một máy đơn, một cụm máy tính phân cụm hay thậm chí là một hệ thống khác có mô hình như đề xuất. Để minh chứng tính khả thi và đúng đắn của mô hình chúng tôi đưa ra bài toán so khớp tài liệu, sử dụng giải thuật PLSA [12]. Bài toán được mô tả như sau: Kho tài liệu gồm các nút lưu trữ, các nút lưu trữ được cấu hình theo mô hình đã đề xuất. Nút lưu trữ nằm phân tán và giao tiếp qua giao thức TCP/IP. Tài liệu có các định dạng .txt, .doc và .pdf. Đầu vào của bài toán là một tệp tin có một trong các định dạng sau: .txt, .doc, hoặc .pdf. Hệ thống dựa trên kho tài liệu để phân tích xem tài liệu đưa vào có những phần nào đoạn nào được sao chép từ các tài liệu khác và kết luận bao nhiêu phần trăm tài liệu là sao chép từ các tài liệu khác. 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, tổ chức ảo, so khớp tài liệu, bảo mật, PLSA**

I. GIỚI THIỆU CHUNG

Vấn đề quản lý tài liệu điện tử nằm phân tán trên mạng 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 vòng năm năm trở lại đây, tính toán lưới [1] nói chung và lưới dữ liệu [3] 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. Thêm vào đó, những ứng dụng yêu cầu tính phân tán về mặt địa lý, chia sẻ dữ liệu phân tán như các hệ thống cảnh báo sóng thần, hoặc dự báo thời tiết… đòi hỏi phải có một cơ sở hạ tầng tính toán

hiệu quả, phù 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.

Với việc phạm vi tính toán lưới trải rộng về mặt địa lý cũng như bao gồm rất nhiều các tổ chức, các trung tâm tính toán khác nhau, thì chứng thực người dùng và ủy quyền trở thành một trong những vấn đề quan trọng nhất của công nghệ tính toán lưới. Các cơ chế phân quyền đơn giản, dựa trên việc đăng ký trực tiếp của người dùng với tài nguyên là không đủ để thực hiện trong các môi trường tính toán lớn và phức tạp. Một mô hình ủy quyền khác được đưa ra dựa trên giấy phép ủy quyền thành viên nhóm [3]. Trong mô hình này, người dùng trên lưới được tổ chức thành các nhóm gọi là tổ chức ảo – Virtual Organization (VO), với cùng phạm vi & chính sách chia sẻ tài nguyên. Mô hình này tích hợp đầy đủ các cơ chế quản lý truy cập tài nguyên cho cả phía người dùng và phía cung cấp tài nguyên, đồng thời làm đơn giản hóa quá trình xây dựng cơ chế quản lý phân quyền người dùng trên lưới.

Việc áp dụng công nghệ tính toán lưới, mà cụ thể là lưới dữ liệu, cùng với các cơ chế ủy quyền dựa trên mô hình Tổ chức ảo cho ta một lời giải phù hợp cho bài toán quản lý tài liệu điện tử. Trong bài báo này, chúng tôi trình bày Giải pháp đề xuất để quản lý tài liệu điện tử nằm phân tán trong các tổ chức ảo, đồng thời đưa ra một số kết quả thử nghiệm và đánh giá việc áp dụng giải pháp trên trong một trường hợp cụ thể và khá điển hình cho vấn đề quản lý tài liệu – bài toán so khớp văn bản.

II. MỘT SỐ VẤN ĐỀ LIÊN QUAN

Vấn đề cộng tác và chia sẻ tài nguyên động trong các tổ chức ảo & giữa các tổ chức ảo với nhau đã được đề cập đến từ lâu, và đã có rất nhiều các giải pháp sử dụng các công nghệ tính toán phân tán khác lưới để giải quyết vấn đề trên. Tuy nhiên các cách tiếp cận này chưa xây dựng được một giải pháp đầy đủ cho vấn đề chia sẻ tài nguyên động, đáp ứng yêu cầu cho mô hình các tổ chức ảo.

Công nghệ Internet & tính toán ngang hàng (Peer-to-peer Computing) (được triển khai trong các dự án như Napster, Freenet hay SETI@home...) thường tập trung vào các giải pháp tích hợp theo chiều dọc cho các vấn đề chuyên biệt, thay vì định nghĩa một giao thức chung cho phép chia sẻ tài nguyên và làm việc cộng tác. Hơn nữa, hình thức chia sẻ cũng còn nhiều hạn chế, như chia sẻ tệp tin không có kiểm soát truy nhập, hay chia sẻ tính toán với một máy chủ tập trung. Khi các ứng dụng trở nên phức tạp hơn và yêu cầu làm việc cộng tác trở nên rõ ràng hơn, thì công nghệ Internet và tính toán ngang hàng cho thấy nhiều hạn chế so với tính toán lưới. Ví dụ, khả năng truy cập một lần (single sign-on), ủy quyền và các công nghệ chứng thực khác cung cấp bởi công nghệ lưới trở nên rất quan trọng và cần thiết khi các dịch vụ chia sẻ dữ liệu và tính toán phải cộng tác với nhau, hay khi chính sách điều khiển truy nhập đến các tài nguyên trở nên phức tạp.

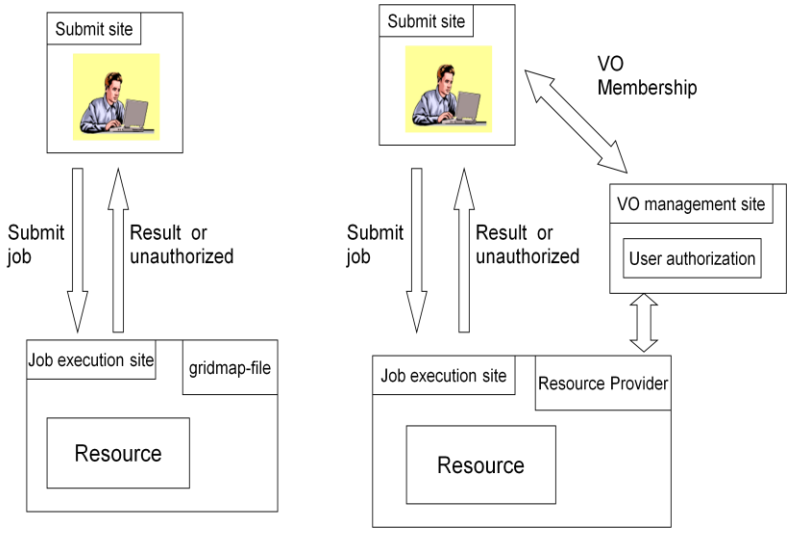
Các công nghệ điện toán doanh nghiệp như CORBA, Enterprise JavaBeans, Java 2 Enterprise Edition, hay DCOM đều được thiết kế để phát triển các ứng dụng phân tán. Các công nghệ này cung cấp các giao diện truy cập tài nguyên chuẩn, các cơ chế triệu gọi phương thức từ xa cũng như các cơ chế giúp đơn giản hóa việc trao đổi tài nguyên trong phạm vi một tổ chức đơn. Tuy nhiên, các cơ chế này lại không giải quyết được các yêu cầu đặt ra với các tổ chức ảo. Việc chia sẻ thường mang tính chất tĩnh & gói gọn trong phạm vi một tổ chức. Dạng thức chủ yếu của các liên tác là theo mô hình Client-server, thay vì là sự phối hợp hoạt động của nhiều tài nguyên.

Các nhà cung cấp dịch vụ ứng dụng (ASPs), các nhà cung cấp dịch vụ lưu trữ (SSPs) và các công ty cung cấp các dịch vụ hosting khác thường triển khai các ứng dụng doanh nghiệp cũng như cung cấp khả năng lưu trữ cho các khách hàng của mình. Công nghệ VPN thường được sử dụng để mở rộng hệ thống mạng của khách hàng, bổ sung thêm các tài nguyên được quản lý và điều khiển bởi các ASPs hay SSPs. Một số SSPs cung cấp các dịch vụ chia sẻ file, truy cập thông qua HTTP, FTP, hay WebDAV với cơ chế mật khẩu hay danh sách điều khiển truy nhập. Chính việc sử dụng VPN hay các cơ chế cấu hình tĩnh này khiến cho việc thực hiện hóa các mô hình chia sẻ tài nguyên giữa các tổ chức ảo trở nên rất khó khăn.

III. QUẢN LÝ TỔ CHỨC ẢO TRÊN MÔI TRƯỜNG LƯỚI

Hiện nay, để thuận tiện quản lý người dùng và truy cập tài nguyên, người dùng trên lưới thường được tổ chức thành các nhóm gọi là tổ chức ảo – Virtual Organization (VO). Tổ chức ảo là tập hợp các cá nhân, đơn vị có cùng phạm vi và chính sách chia sẻ tài nguyên. Các thành viên trong cùng VO cùng nhau chia sẻ tài nguyên lưới trong phạm vi mà bên cung cấp tài nguyên – Resource Provider (RP) quy định cho VO đó.

Ở các tổ chức lưới lớn, các quy chế phân quyền giản đơn, dựa trên sự đăng ký sử dụng trực tiếp của người dùng trên máy chứa tài nguyên là không đủ để thực hiện trong các môi trường tính toán phức tạp. Cụ thể, người dùng lưới sẽ được ánh xạ trực tiếp lên các tài khoản cục bộ nằm trên máy thực thi. Họ sẽ phải liên hệ trực tiếp với các máy này để đăng ký cũng như thay đổi quyền truy cập tài nguyên của mình. Khi quy mô lưới tăng, số lượng người dùng sẽ tăng phi mã theo. Việc các máy cung cấp tài nguyên phải xử lý một số lượng lớn các giao dịch đăng ký sẽ làm giảm hiệu năng của toàn bộ lưới. Để giải quyết vấn đề này, dịch vụ thành viên tổ chức ảo VOMS [4] và dịch vụ đăng ký VOM -VOMRS [5] đã được phát triển, nâng mức quản lý người dùng lên 1 cấp (Hình 1). Việc quản lý người dùng sẽ do VO và các dịch vụ của nó phụ trách. RP sẽ chuyên trách thực hiện các tác vụ lưới.



Hình 1 – Mô hình quản lý người dùng với VOs

VOMS là nỗ lực chung của European DataGrid (CERN) [6] và DataTAG (INFN) [7] nhằm tạo ra một hệ thống cho phép quản lý thông tin phân quyền của người dùng trong VO theo thời gian thực. Các thông tin phân quyền được lưu bao gồm các thông tin chứng thực, vai trò người dùng và các quyền của họ trong VO.

Cụ thể, VO được tổ chức phân cấp. Người dùng được chia vào các nhóm tùy theo tổ chức họ tham gia và các nhiệm vụ họ cần làm. Mỗi nhóm sẽ có người quản trị phụ trách quản lý, cấp quyền cho các thành viên. Tùy theo yêu cầu công việc, sẽ quyết định có nên tạo các nhóm con mới trực thuộc.

Người dùng trong nhóm được đặc trưng bởi vai trò và các quyền tương ứng với vai trò đó. Như vậy, người dùng trong VO được đặc trưng bởi 3 thuộc tính cơ bản là nhóm, vai trò và quyền. Cả 3 thuộc tính này được gọi chung là "Tên thuộc tính đầy đủ" - "Fully Qualified Attribute Names" (FQAN).

Cú pháp:

/VO[/group[/subgroup(s)]][/Role = role][/Capability = cap]

Ví dụ: Vai trò Administrator trong group AGP của VO hpcc.hut.edu.vn là:

/hpcc.hut.edu.vn/AGP/Role=Administrator

Tuy vậy, quy trình đăng ký người dùng vào VO của VOMS vẫn còn giản đơn, ví dự như không yêu cầu người dùng ký vào biên bản thỏa thuận sử dụng lưới AUP, không hỗ trợ quản lý hiệu lực quyền thành viên… Để khắc phục những hạn chế này, VOMS eXtension (VOX) được phát triển. Kết quả đến cuối 2004, dịch vụ quản lý đăng ký tổ chức ảo VOMRS [5] đã ra đời.

VOMRS đã giải quyết được những hạn chế của VOMS như bắt buộc người dùng phải đồng ý vào cam kết sử dụng lưới trước khi gia nhập VO, cho phép người dùng được chủ động gửi yêu cầu tham gia nhóm, bổ sung thêm tính năng tạm treo quyền thành viên, định kỳ kiểm tra thời gian quyền thành viên còn hiệu lực cho người quản trị. VOMRS phối hợp đồng bộ với VOMS tạo ra hệ thống hoàn hảo tích hợp được sức mạnh của bản thân 2 dịch vụ.

Hiện tại, VOMS & VOMRS đang được sử dụng rất nhiều trong các tổ chức lưới lớn trên khắp thế giới như PRAGMA, EGEE, USCMS, LHC... Trên PRAGMA, VO được phân thành các lưới con như Avian-Flu-Grid, e-AIRS, Geo, NIMROD... Các lưới con này, chẳng hạn như Geo Grid [8] lại được phân tiếp thành các nhóm nghiên cứu về Disaster, Geology, Environment... Quá trình phân nhóm tiếp tục cho đến khi phù hợp với yêu cầu phân quyền và sử dụng tài nguyên trong nhóm.

Trong các nhóm, sẽ có nhiều vai trò như Member, Site Administrator, Representative, Group Manager... Phân vai trò đảm bảo cho các thành viên chỉ được sử dụng tài nguyên lưới trong phạm vi quy định.

Mô hình hệ thống quản lý tài liệu điện tử trên môi trường lưới có sự kết hợp của nhiều tổ chức, với sự tham gia của nhiều loại người dùng nên việc ứng dụng mô hình tổ chức ảo vào hệ thống, làm đơn giản việc quản lý người dùng trên môi trường lưới là rất có ý nghĩa. Người dùng hệ thống được phân nhóm theo các lĩnh vực quan tâm (như Công nghệ thông tin, Toán học, Lý học, Hóa học...). Các vai trò được chia ra là Nhà nghiên cứu và người dùng thường. Các dịch vụ sẽ được phân quyền là tìm kiếm tài liệu, upload tài liệu và so khớp tài liệu. Khi đó, bảng phân quyền sẽ là:

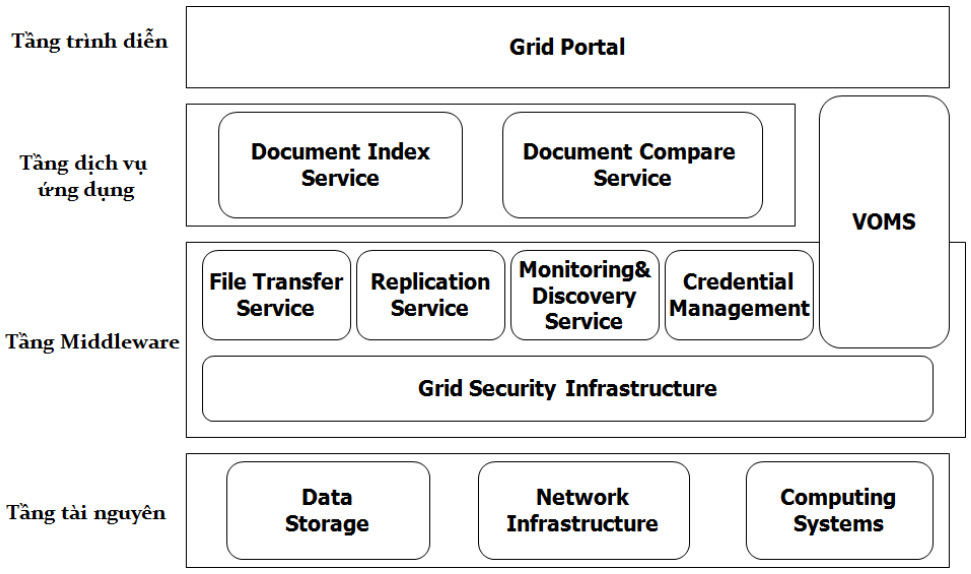
Người dùng	So khớp tài liệu	Tìm kiếm tài liệu	Upload tài liệu
Nhà nghiên cứu	X	X	X
Người dùng thường	O	X	O

X: Được phép

O: Không được phép

IV. GIẢI PHÁP QUẢN LÝ TÀI LIỆU ĐIỆN TỬ PHÂN TÁN TRONG TỔ CHỨC ẢO

Giải pháp đưa ra kết nối các tài nguyên tính toán, lưu trữ của các tổ chức, trung tâm tính toán thành một lưới dữ liệu, mỗi tổ chức là một nút lưới. Người dùng truy cập và sử dụng các chức năng của hệ thống thông qua một cổng thông tin. Ngoài cơ chế xác thực người dùng qua portal, các thông tin phân quyền bao gồm các thông tin chứng thực, vai trò người dùng và các quyền của họ trong VO được xác nhận thông qua máy chủ dịch vụ thành viên tổ chức ảo - Virtual Organization Membership Service (VOMS). Từ đó, người dùng có thể sử dụng được các dịch vụ của hệ thống với quyền hạn của mình. Hình 2 minh họa Mô hình kiến trúc của hệ thống



Hình 2- Mô hình kiến trúc hệ thống

- 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 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 [9] là một trong những grid middleware được sử dụng rộng rãi nhất trong các dự án tính toán lưới.

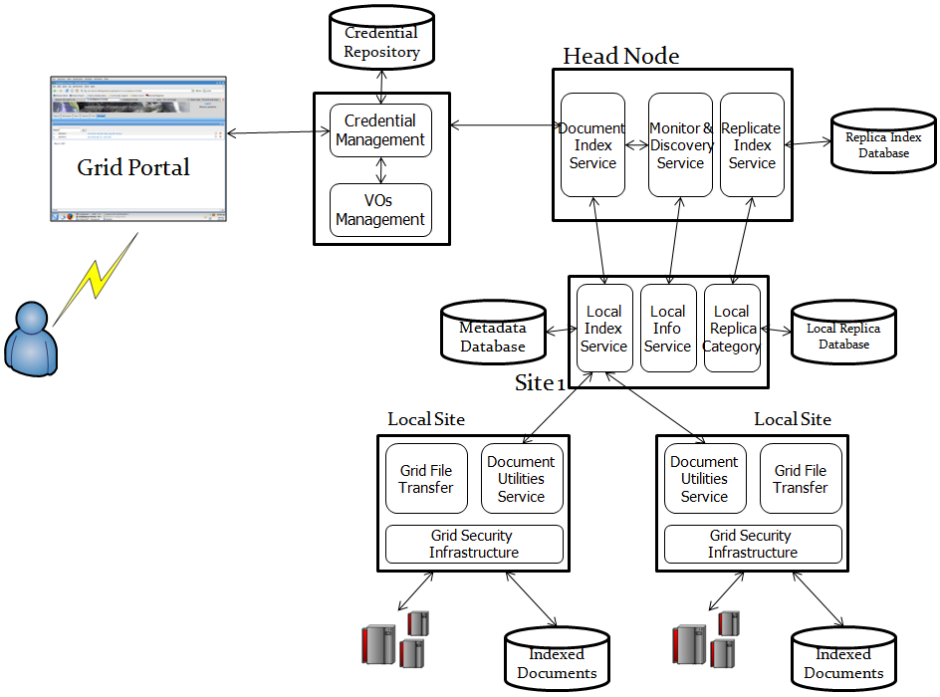
  Một số dịch vụ được cung cấp bởi tầng Middleware như:

  - Hạ tầng bảo mật lưới
  - Dịch vụ truyền file
  - Dịch vụ quản lý bản sao
  - Dịch vụ thông tin hệ thống

- Dịch vụ quản lý giấy phép ủy quyền

- Tầng dịch vụ ứng dụng: dựa trên cơ sở hạn tầng lưới (các dịch vụ nền, các APIs được cung cấp bởi tầng middleware), tầng dịch vụ ứng dụng cung cấp các dịch vụ lưới mà các nhà phát triển có thể triển khai, 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) [10], 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 3 minh họa Mô hình triển khai của hệ thống:



Hình 3 – Mô hình triển khai hệ thống

Người dùng truy cập hệ thống thông qua grid portal. Người dùng portal sẽ được ánh xạ đến người dùng lưới thông qua dịch vụ quản lý giấy phép ủy quyền (Credential Management Service). Dịch vụ này quản lý tất cả các giấy phép ủy quyền của người dùng lưới bao gồm cả các thông tin về tổ chức ảo & nhóm tham gia, cũng như vai trò của người dùng trong tổ chức ảo đó. Các thông tin này được cung cấp bởi VOs Management Service.

Các yêu cầu của người dùng, như tìm kiếm, so khớp văn bản, cùng với các thông tin về tổ chức ảo được gửi đến Head Node xử lý. Dựa trên thông tin về tổ chức ảo, yêu cầu của người dùng sẽ được gửi đến các trạm vật lý (Sites) tương ứng tham gia vào tổ chức ảo đó để xử lý. Bản thân các trạm lại được tổ chức phân cấp thành các trạm cục bộ (Local Sites), trên mỗi trạm cục bộ này triển khai các dịch vụ tiện ích (Ultilities Service) để xử lý các yêu cầu trên tập các dữ liệu được lưu trữ cục bộ của mình. Việc tổ chức các trạm cục bộ là tùy thuộc chính sách vào mỗi tổ chức. Thông tin mô tả về các dịch vụ và dữ liệu được lưu trữ phân tán trên các trạm cục bộ được lưu trong một Cơ sở dữ liệu mô tả (Metadata Database). Những thông tin này được các Trạm sử dụng để phân tích & gửi các yêu cầu đến các trạm cục bộ tương ứng để xử lý.

Bên cạnh đó, để đảm bảo nhất quán về dữ liệu, hệ thống triển khai một cơ chế quản lý bản sao phân cấp với các Cơ sở dữ liệu bản sao cục bộ trên các Trạm (Local Replica Database) và một Cơ sở dữ liệu quản lý bản sao toàn cục (Replica Index Service) trên Trạm trung tâm (Head node). Thông tin trạng thái tài nguyên của hệ thống cũng được đảm bảo cập nhật liên tục & thông suốt với cơ chế quản lý thông tin hệ thống phân cấp. Các dịch vụ quản lý thông tin cục bộ (Local Information Service) quản lý thông tin tài nguyên trong phạm vi trạm của mình & cập nhật thông tin lên dịch vụ Giám sát & quản lý thông tin toàn hệ thống (Monitoring & Discovery Service).

Toàn bộ hệ thống được triển khai trên môi trường lưới, các dịch vụ tiện ích được cài đặt theo chuẩn dịch vụ lưới [11] và được triển khai trên kho dịch vụ lưới. Việc triển khai các dịch vụ phân cấp trên các trạm cục bộ cũng như phát triển các dịch vụ theo chuẩn dịch vụ lưới giúp cho khả năng mở rộng hệ thống trên các trạm là rất dễ dàng.

V. THỬ NGHIỆM MÔ HÌNH VỚI BÀI TOÁN SO KHỚP TÀI LIỆU

Với việc áp dụng mô hình tổ chức ảo, phân chia người dùng thành các nhóm nghiên cứu theo lĩnh vực (gọi là các tổ chức ảo), các tài liệu điện tử cũng được lưu trữ & xử lý theo các lĩnh vực riêng biệt, thì việc áp dụng mô hình đề xuất để giải quyết các vấn đề trong quản lý tài liệu điện tử là rất phù hợp.

Để đánh giá được hiệu quả của việc áp dụng mô hình đề xuất vào giải quyết một vấn đề cụ thể trong quản lý tài liệu điện tử, chúng tôi đưa ra mô hình thử nghiệm là Bài toán so khớp tài liệu trên môi trường phân tán.

Phát biểu bài toán: Với mỗi tài liệu đầu vào, xác định mức độ trùng khớp (về ngữ nghĩa) so với các tài liệu trong kho lưu trữ hiện có

Tính chất: Kho tài liệu thuộc nhiều lĩnh vực, nằm phân tán trên nhiều site, mỗi site có các chính sách truy cập & quản lý người dùng riêng.

Giải thuật so khớp được sử dụng: giải thuật PLSA (Probabilistic Latent Semantic Analysis) [12]. PLSA là phương pháp phân tích nội dung document theo hướng tiếp cận ngữ nghĩa, sử dụng các công cụ xác suất thống kê. Đầu vào của PLSA là tập các văn bản & ma trận xuất hiện các từ trong văn bản,sử dụng các phương pháp xác suất thống kê tìm ra tập các chủ đề ẩn trong văn bản, từ đó biểu diễn văn bản là một vector đặc trưng cho văn bản đó.

Cũng như các giải thuật so khớp hay tìm kiếm dựa trên ngữ nghía khác, PLSA cho kết quả tốt hơn rất nhiều khi kho tài liệu mẫu được phân loại theo nhóm các lĩnh vực (phân loại theo nội dung). Một số ưu điểm của PLSA so với các giải thuật so khớp văn bản khác đó là:

- Khắc phục được vấn đề synonym (từ đồng nghĩa) và polysemy (từ đa nghĩa)
- Độ phức tạp của thuật toán thấp (O(mn) với m: số document, n: số word)
- Bài toán có thể phân rã (áp dụng trên môi trường hệ phân tán)

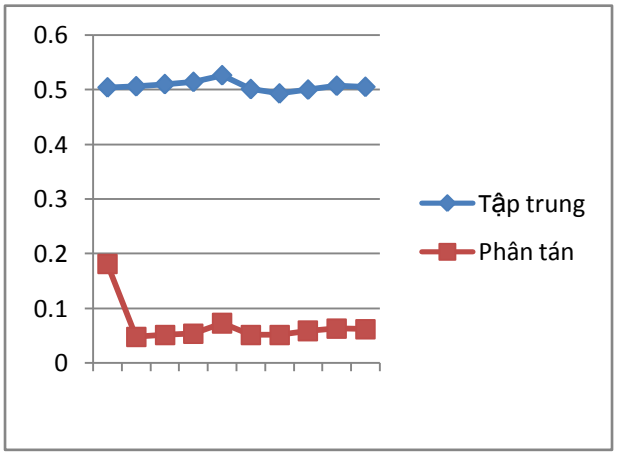
Mô hình thử nghiệm:

- Xây dựng lưới dữ liệu cho hệ thống, mỗi trường đại học, tổ chức nghiên cứu là một nút lưới. Trên mỗi nút lưới cài đặt các dịch vụ tìm kiếm, upload & so khớp tài liệu. Kho tài liệu trên mỗi nút lưới được phân loại thành các nhóm lĩnh vực & được lưu trữ trên các trạm cục bộ của các nút lưới đó.
- Sử sụng mô hình VO, phân nhóm người dùng theo các lĩnh vực & vai trò:

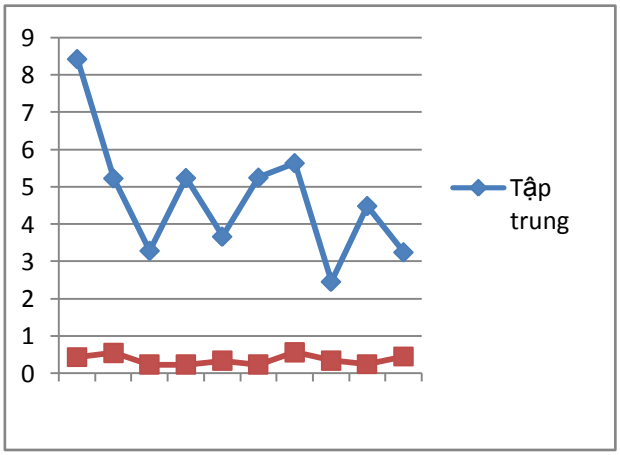
- Nhóm lĩnh vực (Groups): IT, Literature, Math, Physic, Chemical…
- Vai trò (Roles) trong mỗi nhóm lĩnh vực: Nhà nghiên cứu, người dùng thường.

Khi người dùng đăng nhập vào, hệ thống sẽ xác định người dùng đó thuộc nhóm lĩnh vực nào (thuộc VO nào) & có vai trò gì trong nhóm đó. Từ đó, xác định được các dịch vụ mà người dùng đó có thể sử dụng được cũng như gửi các yêu cầu xử lý (tìm kiếm, so khớp) đến các trạm nằm trong nhóm lĩnh vực tương ứng của người dùng đó.

Kết quả thử nghiệm trên hơn 2000 tài liệu, được phân theo 10 nhóm lĩnh vực khác nhau và được so sánh với việc xử lý tập trung (không phân nhóm tài liệu & không áp dụng mô hình Tổ chức ảo vào phân nhóm người dùng theo các lĩnh vực).



Biểu đồ 1 – Thời gian xử lý (giây)



Biểu đồ 2 – Tỷ lệ lỗi (%)

Qua 2 biểu đồ kết quả trên, có thể thấy thời gian xử lý được cải thiện đáng kể, do các văn bản đã được phân loại thành các nhóm lĩnh vực & các yêu cầu của người dùng cũng được phân loại & xử lý trong phạm vi các nhóm nghiên cứu (các tổ chức ảo) mà người đó tham gia. Bên cạnh đó, kết quả xử lý cũng chính xác hơn – được minh họa bởi biểu đồ tỷ lệ lỗi. Tỷ lệ lỗi ở đây được xác định bởi tỷ lệ các lần so khớp các văn bản có kết quả không chính xác (luôn có tỷ lệ sai sót nhất định đối với các giải thuật sử dụng phương pháp tiếp cận theo xác suất thống kê) – được xác định bằng trực quan. Chất lượng xử lý được nâng lên qua việc áp dụng mô hình tổ chức ảo, phân nhóm người dùng theo các lĩnh vực cũng như phân loại các tài liệu theo các lĩnh vực & xử lý các yêu cầu người dùng trên tập các văn bản đã phân loại đó.

VI. KẾT LUẬN

Trong bài báo này, chúng tôi đã đề xuất giải pháp lưới dữ liệu quản lý tài liệu điện tử trong các tổ chức ảo. Đồng thời đưa ra những thử nghiệm và đánh giá hiệu quả của việc áp dụng giải pháp đề xuất vào giải quyết một vấn đề cụ thể trong quản lý tài liệu, đó là so khớp văn bản. Những kết quả thử nghiệm và chạy thử hệ thống cho thấy: hệ thống 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 tổ chức ảo. Qua đó, có thể thấy mô hình giải pháp đề ra là phù hợp để giải quyết các vấn đề quản lý và chia sẻ tài nguyên phân tán trên mạng, tài nguyên thuộc về các tổ chức với những chính sách quản lý truy cập khác nhau, mà bài toán quản lý tài liệu điện tử chỉ là một trường hợp riêng.

TÀI LIỆU THAM KHẢO

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

[2]. Ian Foster, Carl Kesselman, and Steven Tuecke, ***"The Autonomy of the Grid, Enabling Scalable Virtual Organizations,"*** Int. J. Supercomput. Appl. 2002.

[3]. 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.

[4]. 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.

[5]. R. Alfieria, R. Cecchinib, V. Ciaschinic, L. dell'Agnellod, A´ . Frohnere, K. Lo″renteyf , F. Spataro, ***" From gridmap-file to VOMS: managing authorization in a Grid environment "***, 2005.

[6]. A. Ceccanti , V. Ciaschini, M. Dimou, G. Garzoglio, T. Levshina, S. Traylen, V. Venturi, ***"VOMS/VOMRS Utilization patterns and convergence plan"***, 2008.

[7]. European DataGrid Project - eu-datagrid.web.cern.ch

[8]. DataTAG Project - http://datatag.web.cern.ch

[9]. GeoGrid - http://www.geogrid.org

[10]. Globus Toolkit Project - www.globus.org

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

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

[13]. Thomas Hofmann, ***"Probabilistic Latent Semantic Analysis"***, EECS Department, Computer Science Division, University of California, Berkeley and International Computer Science Institute, Berkeley, CA, 1999

[14]. Bart Jacob, Luis Ferreira, Norbert Bieberstein, Candice Gilzean, Jean-Yves Girard, Roman Strachowski, Seong (Steve) Yu, ***"Enabling Applications for Grid Computing with Globus"***, IBM Redbooks, 2003