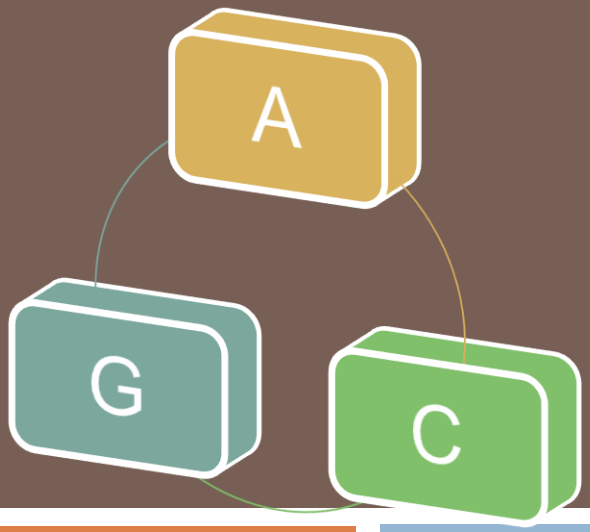
BÁO CÁO SINH VIÊN NGHIÊN CỨU KHOA HỌC

Đề tài :

Hệ thống tìm kiếm và so khớp tài liệu liên trường đại học



GVHD:

PGS. TS Nguyễn Thanh Thuỷ

SVTH:

Tô Trọng Hiến, Nguyễn Hồng Thanh,

Nguyễn Việt Phương, Nguyễn Duy Hoàng,

Nội dung trình bày

- Đặt vấn đề
- Mô hình đề xuất
- Dịch vụ tìm kiếm & so khớp
- Đóng góp và hướng phát triển

Đặt vấn đề

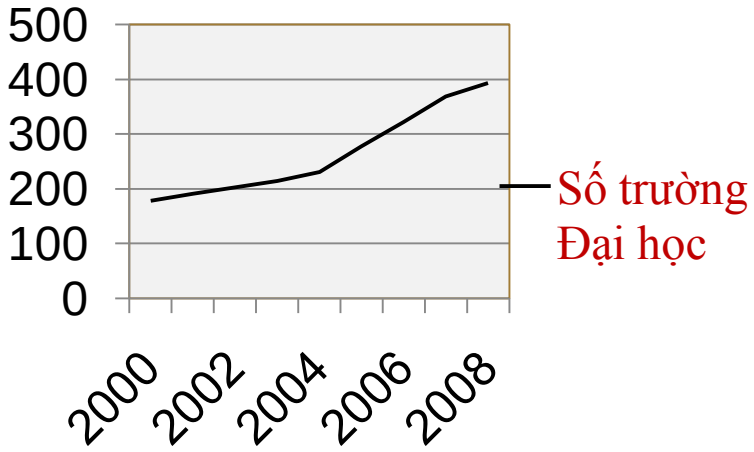
- Nhu cầu một hệ thống quản lý tài liệu, luận văn liên trường đại học là rất lớn

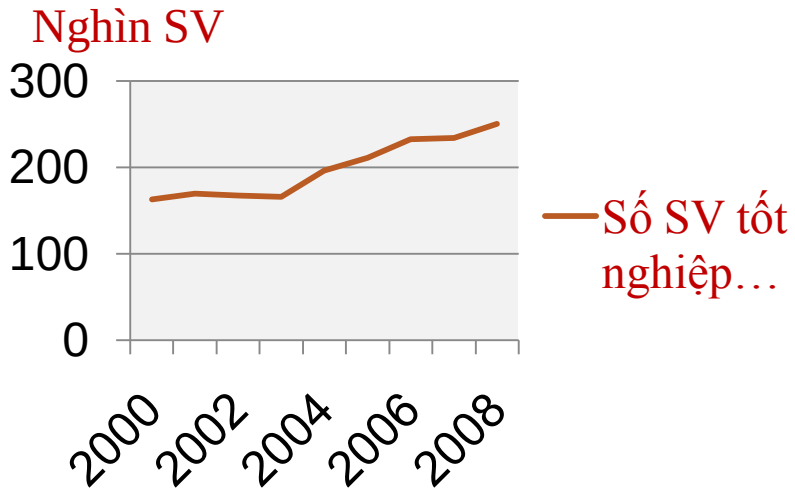
- Các hiện tượng gian lận, sao chép trong học tập xuất hiện ngày một nhiều
  => giảm chất lượng tài liệu, luận văn

Một hệ thống vừa cho phép quản lý vừa có khả năng so khớp tài liệu liên trường

Những khó khăn

- Số lượng tài liệu luận văn rất lớn, lưu trữ phân tán trên các trường đại học





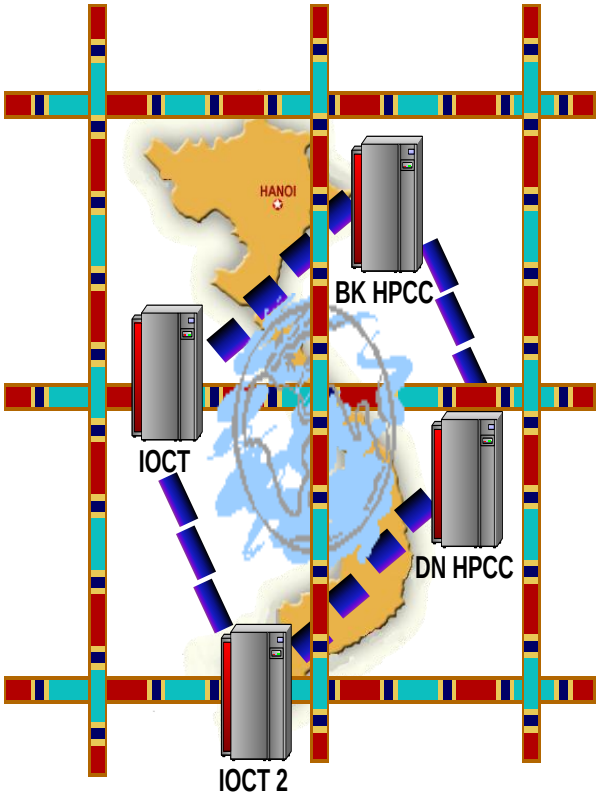
- Chính sách với người dùng & quản lý truy cập là khác nhau
- Các công nghệ phân tán hiện tại còn hạn chế: CORBA và Enterprise Java
- Các phương pháp so khớp cổ điển không đáp ứng được

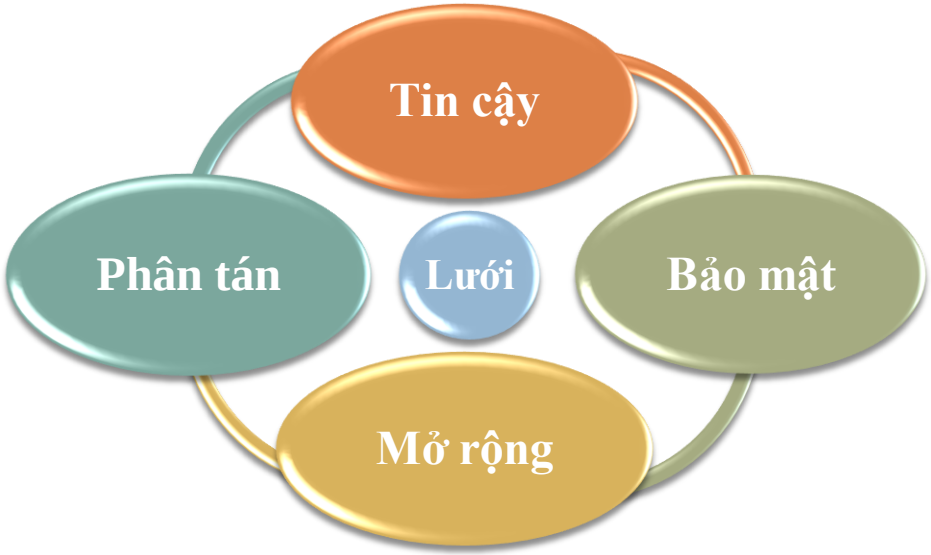
Nội dung trình bày

- Đặt vấn đề
- Mô hình đề xuất
- Dịch vụ tìm kiếm & so khớp
- Đóng góp và hướng phát triển

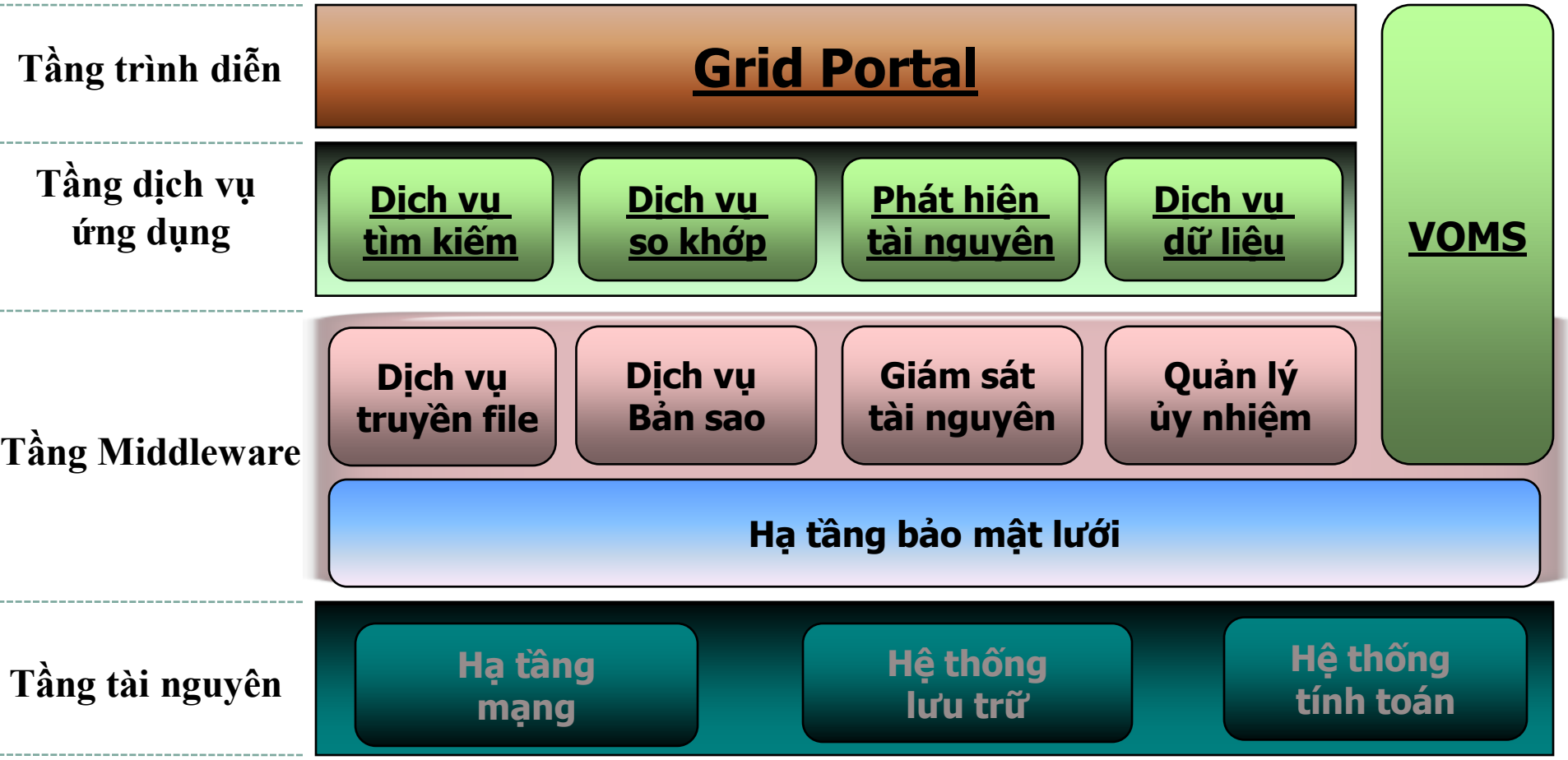
Mô hình đề xuất

- Lưới dữ liệu liên trường đại học, trong đó mỗi trường là một nút lưới
- Việc tìm kiếm & so khớp tài liệu được thực hiện **phân tán** trên các nút
- Cổng thông tin cho phép người dùng dễ dàng tiếp cận hệ thống



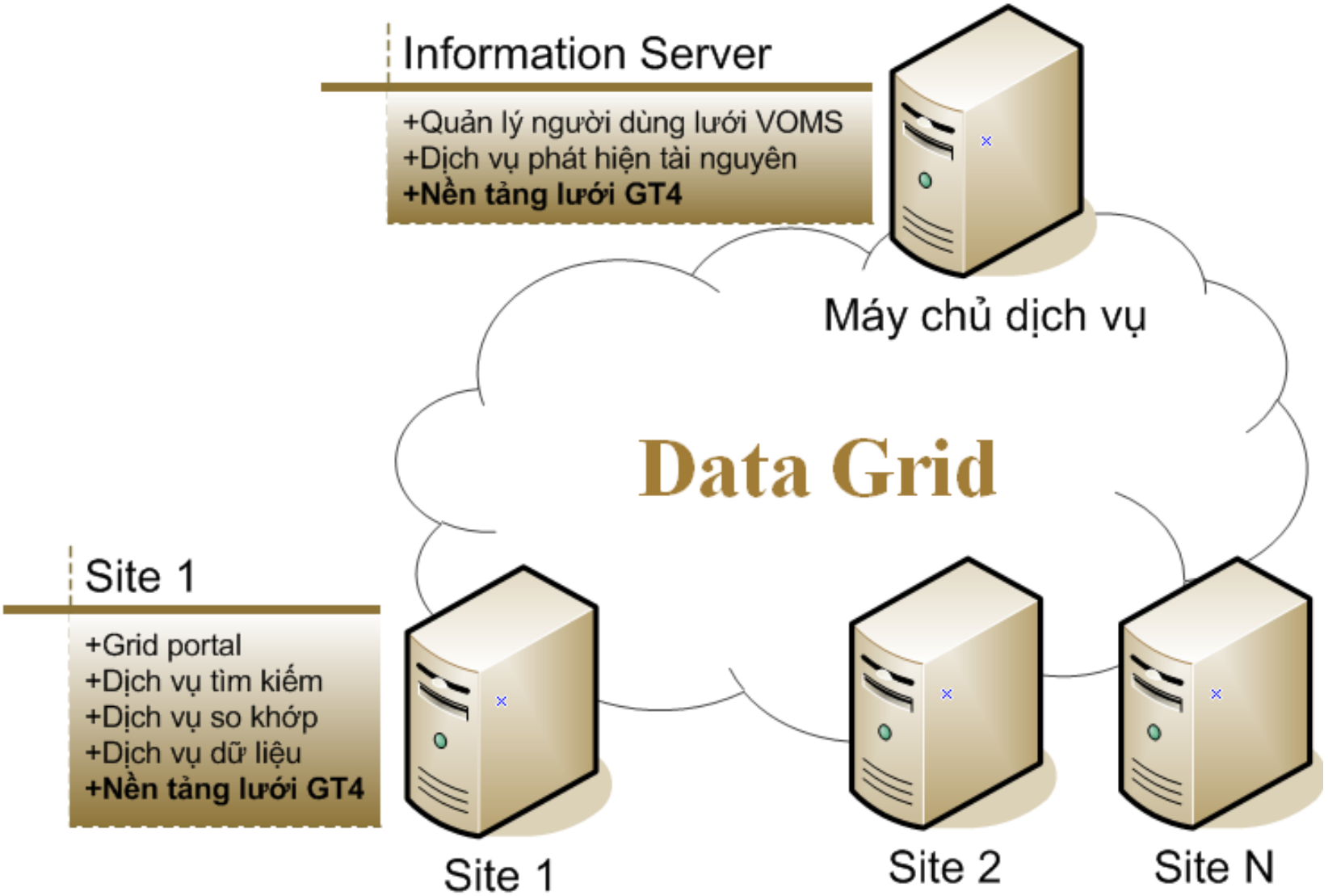


Mô hình kiến trúc hệ thống



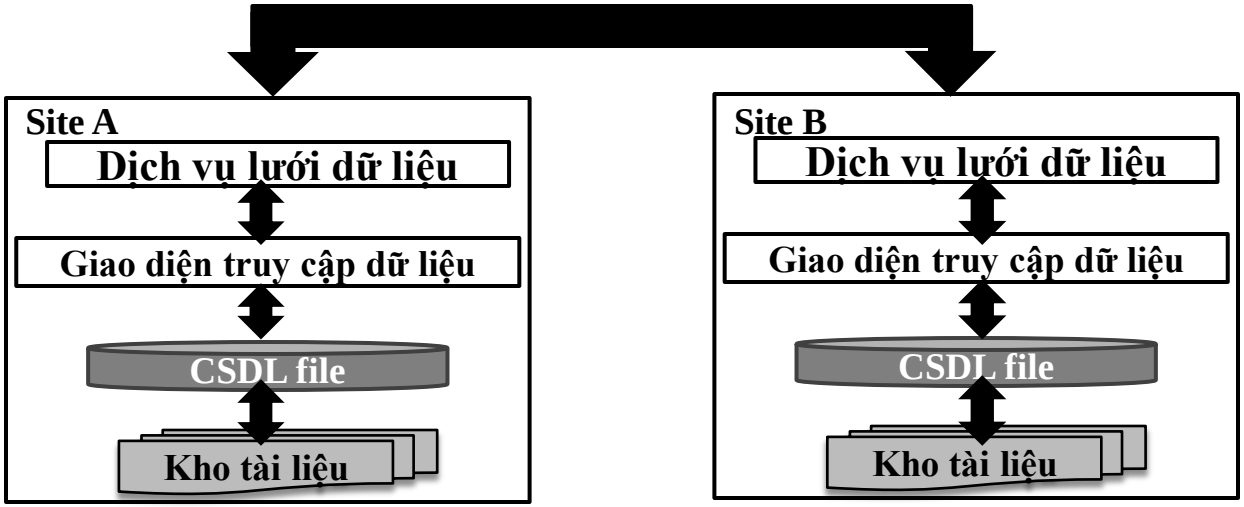
MÔ HÌNH KIẾN TRÚC CÁC TẦNG

Mô hình triển khai



Lưới dữ liệu

- Kết nối dữ liệu phân tán từ các trường
- Trong suốt với người dùng



p truy cập dữ liệu liên trường

- Khả năng tạo lập bản sao

=>Tăng tính tin cậy và hiệu năng

Nội dung trình bày

- Đặt vấn đề
- Mô hình đề xuất
- Dịch vụ tìm kiếm & so khớp
- Đóng góp và hướng phát triển

Tìm kiếm tài liệu phân tán

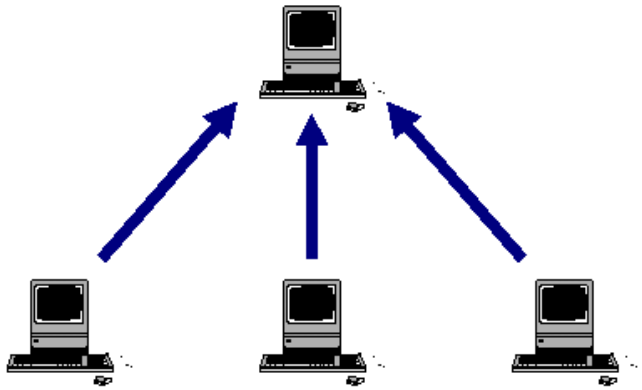
- -

  n

- p Cue-Validity Variance

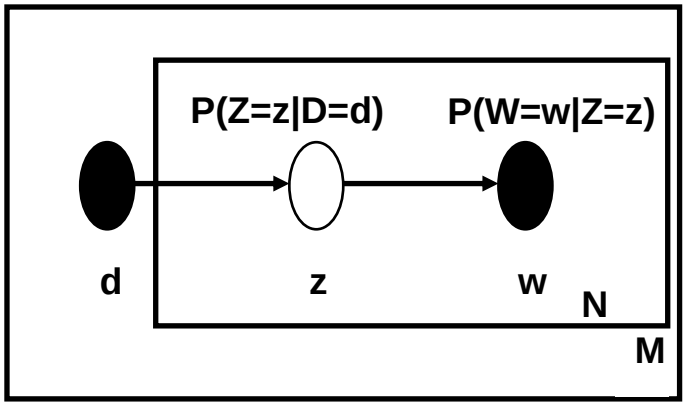
- Ưu điểm:
  - Tốc độ tổng hợp dữ liệu nhanh
  - Lượng dữ liệu trao đổi trong quá trình tổng hợp thấp

  => giảm băng thông hệ thống



Vấn đề so khớp tài liệu

- PLSA (Probabilistic Latent Semantic Analysis): phương pháp phân tích nội dung tài liệu theo hướng tiếp cận ngữ nghĩa
- PLSA xuất phát từ mô hình Aspect (Mô hình biến ẩn)

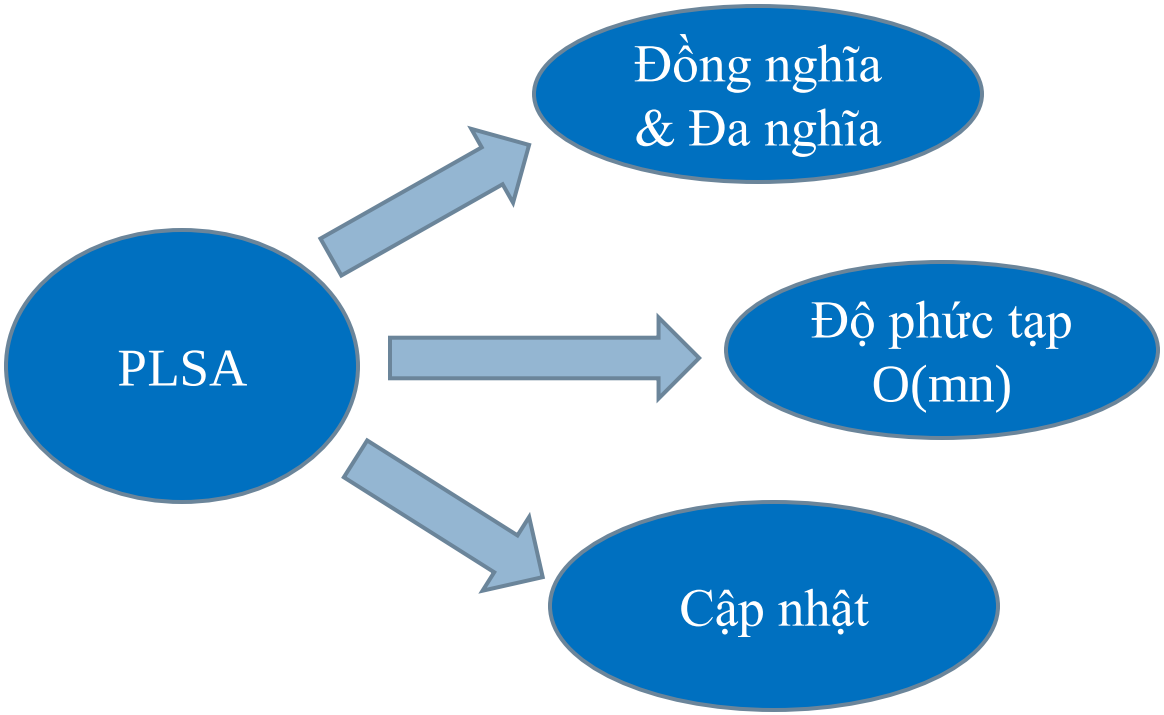


Biểu diễn đồ thị của mô hình Aspect, N: số từ trong tài liệu, M: số tài liệu

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and

Ưu điểm của PLSA

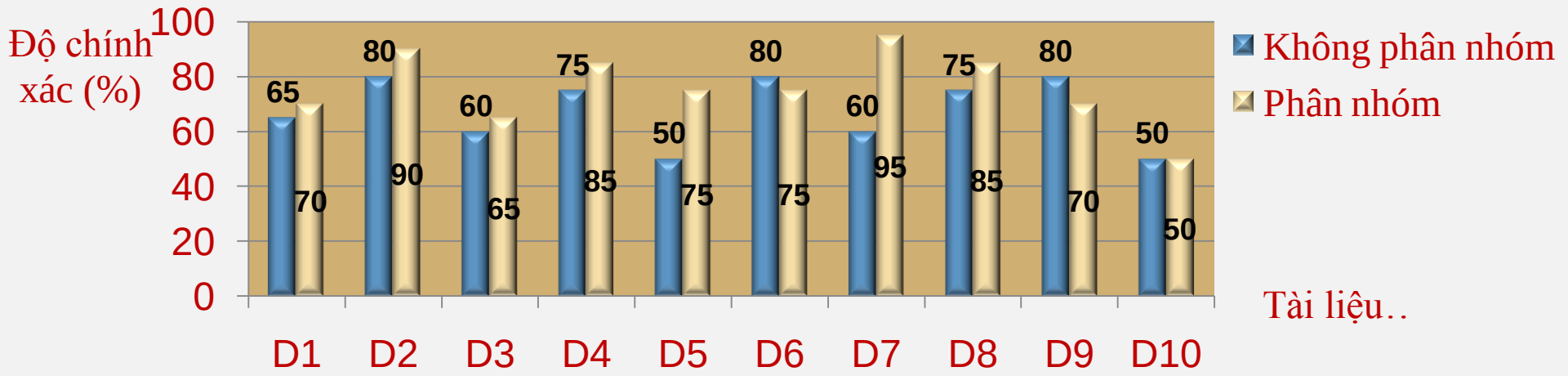


Số chủ đề càng lớn thì độ chính xác càng cao

Thực nghiệm

- 240 abstract của các bài báo khoa học trên ieee.org
- Bộ kết quả chuẩn:
  - Chọn 10 tài liệu chuẩn để truy vấn.
  - Mỗi tài liệu chọn 20 tài liệu có nội dung liên quan đến nó nhất.

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

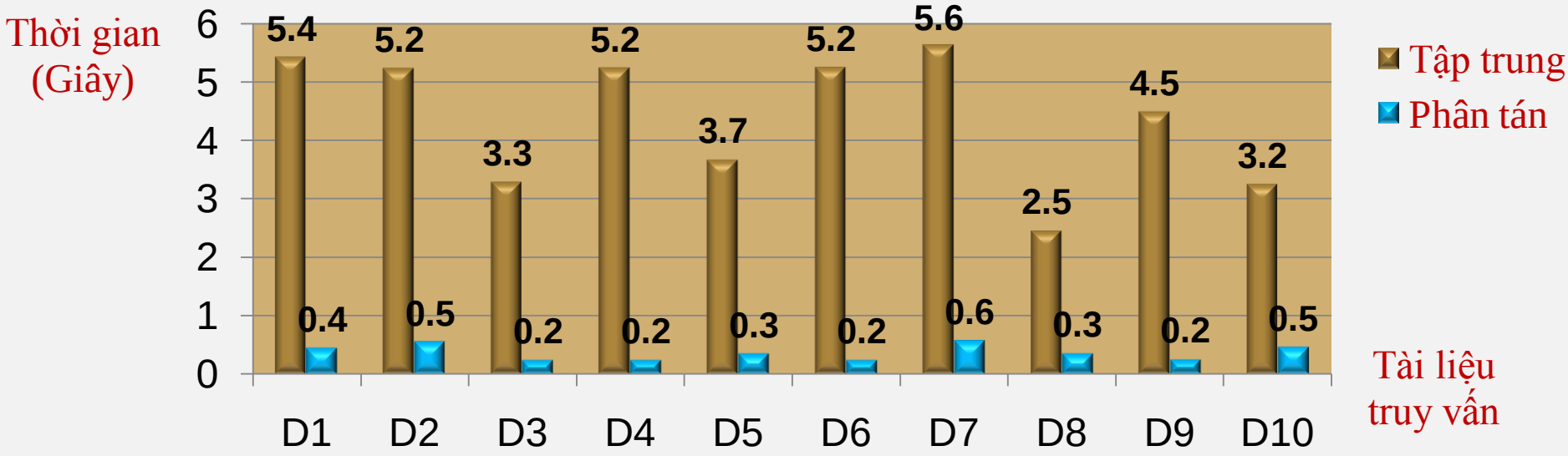


Biểu đồ độ chính xác

Thực nghiệm

- 1000 abstract của các bài báo khoa học trên science direct theo nhiều chủ đề: IR, IR, Grid …

=> Quá trình tìm kiếm và so khớp tài liệu phân tán cho tốc độ nhanh hơn nhiều so với lưu trữ dữ liệu tập trung



Biểu đồ thử nghiệm về tốc độ

Nội dung trình bày

- Đặt vấn đề
- Mô hình đề xuất
- Dịch vụ tìm kiếm & so khớp
- Đóng góp và hướng phát triển

Những đóng góp của đồ án

- Xây dựng hệ thống tìm kiếm và so khớp tài liệu liên trường đại học

Những đóng góp của đồ án

- Module VOMS quản lý người dùng đăng ký lưới

Hướng phát triển

- Hoàn thiện hệ thống quản lý bản sao
- Tiếp tục cải tiến giải thuật so khớp và tìm kiếm cả về mặt tốc độ xử lý lẫn độ chính xác

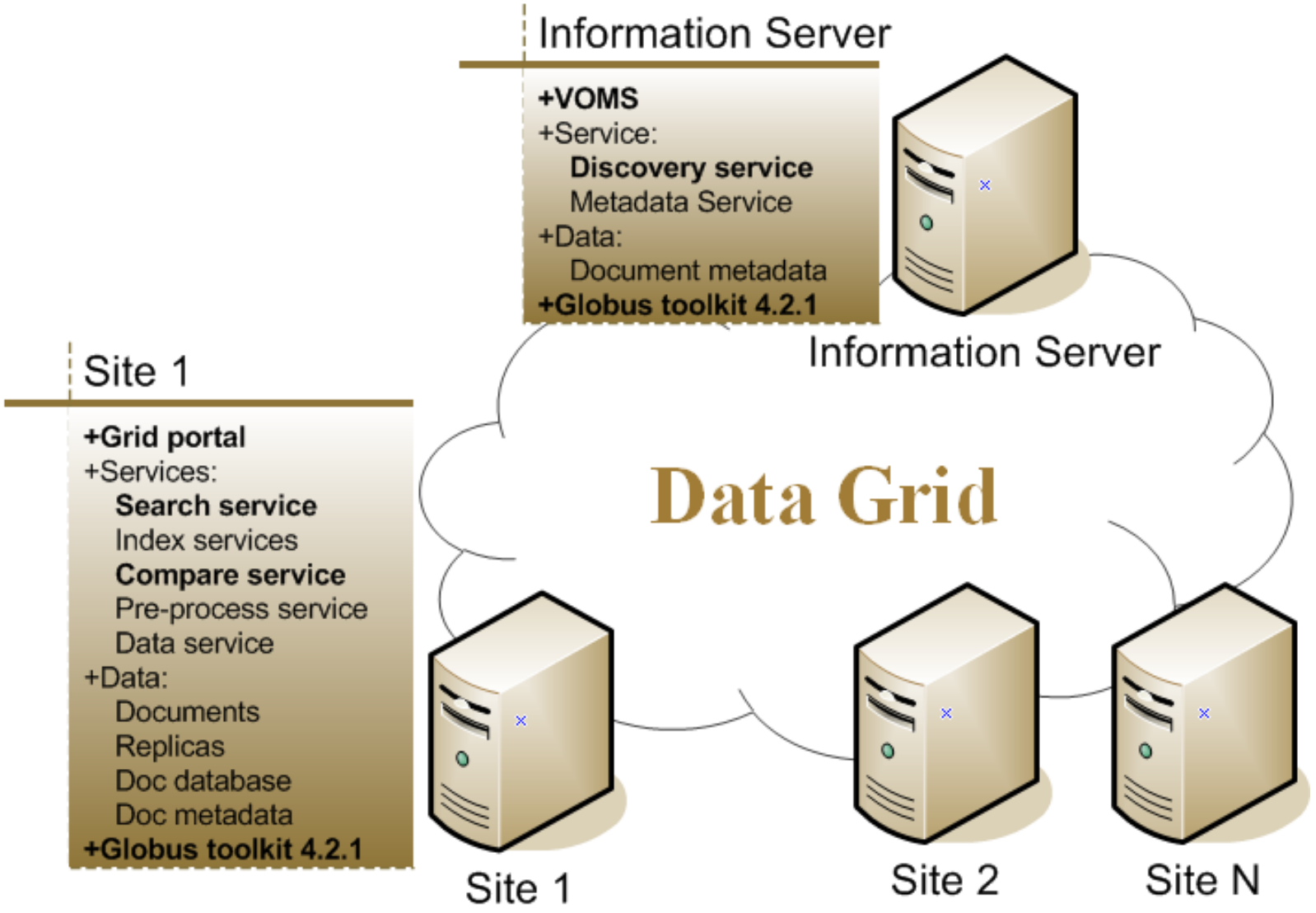
Q&A

Chúng em xin chân thành cảm ơn!

Q&A

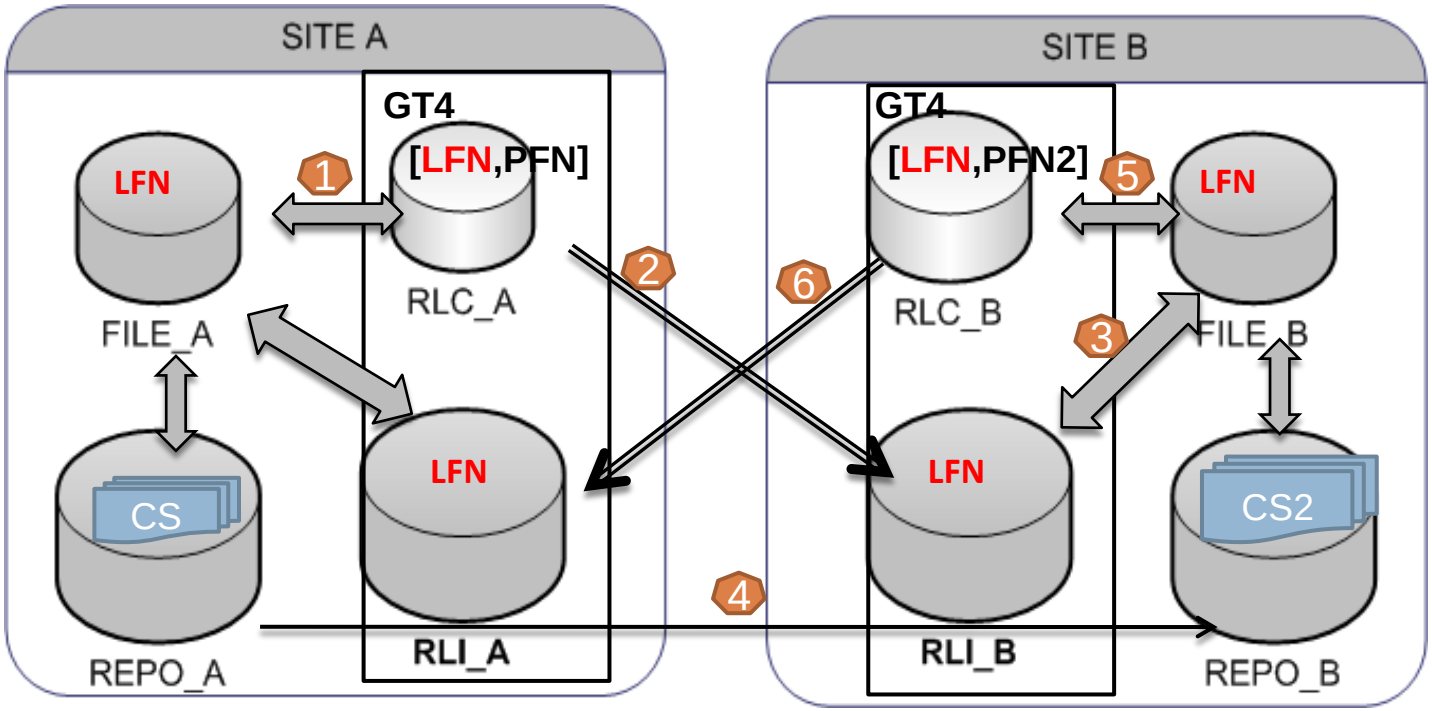
Chúng em xin chân thành cảm ơn!

Mô hình triển khai



Lưới dữ liệu

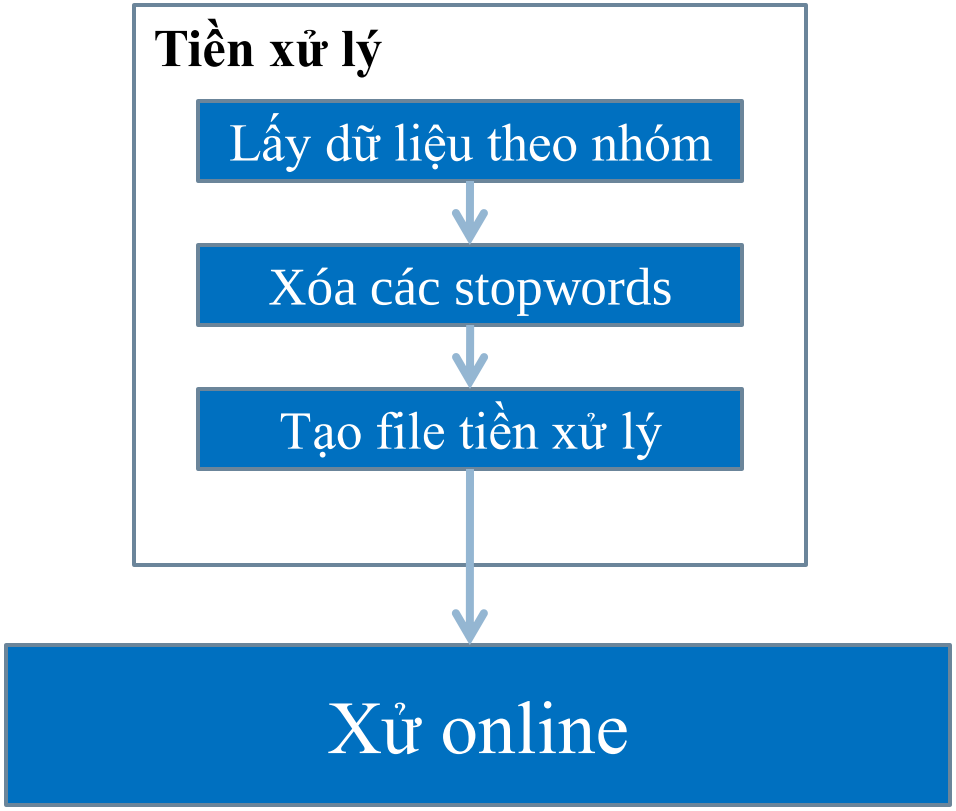
- Khả năng tạo lập bản sao
  - Tăng tính tin cậy và hiệu năng

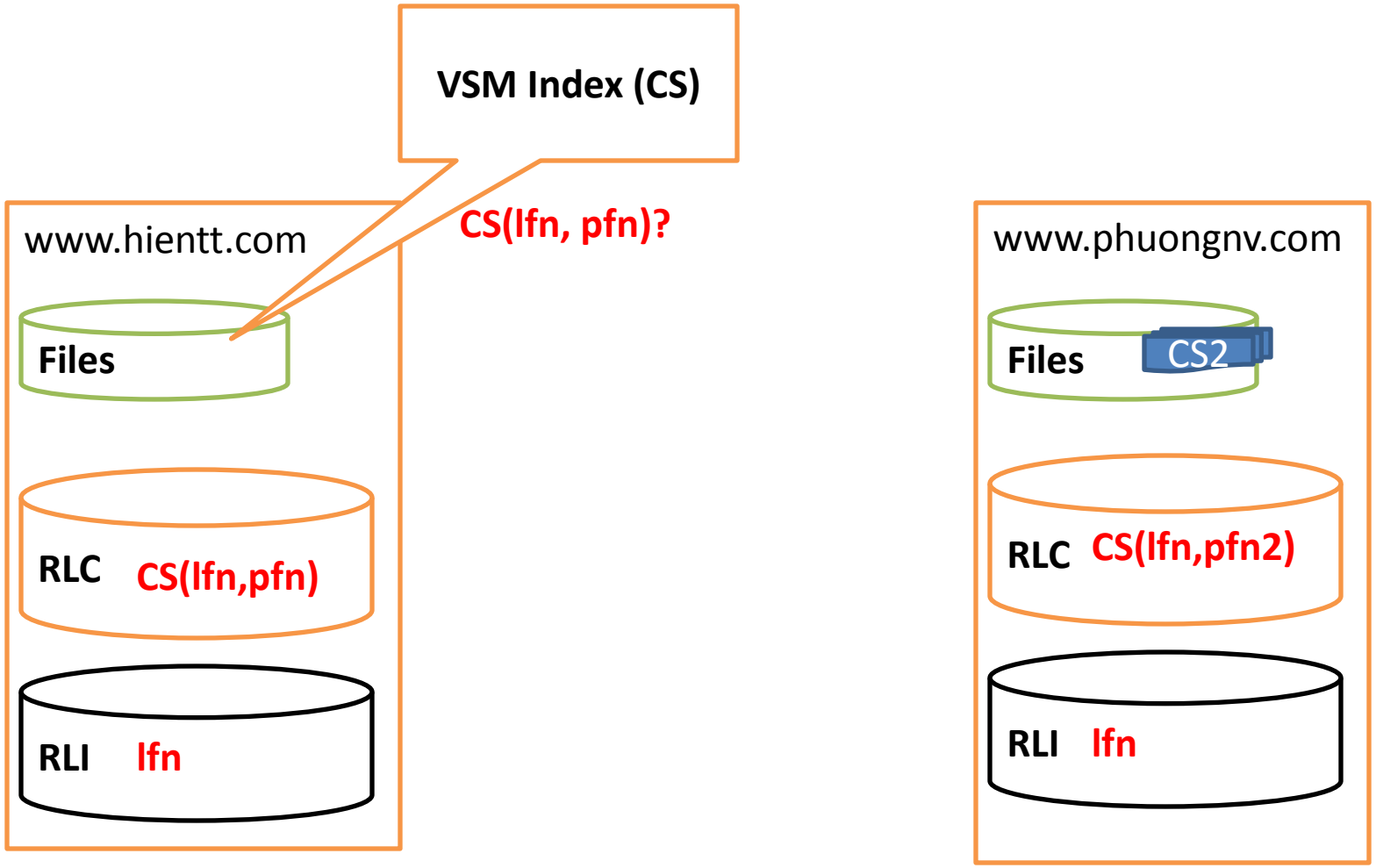


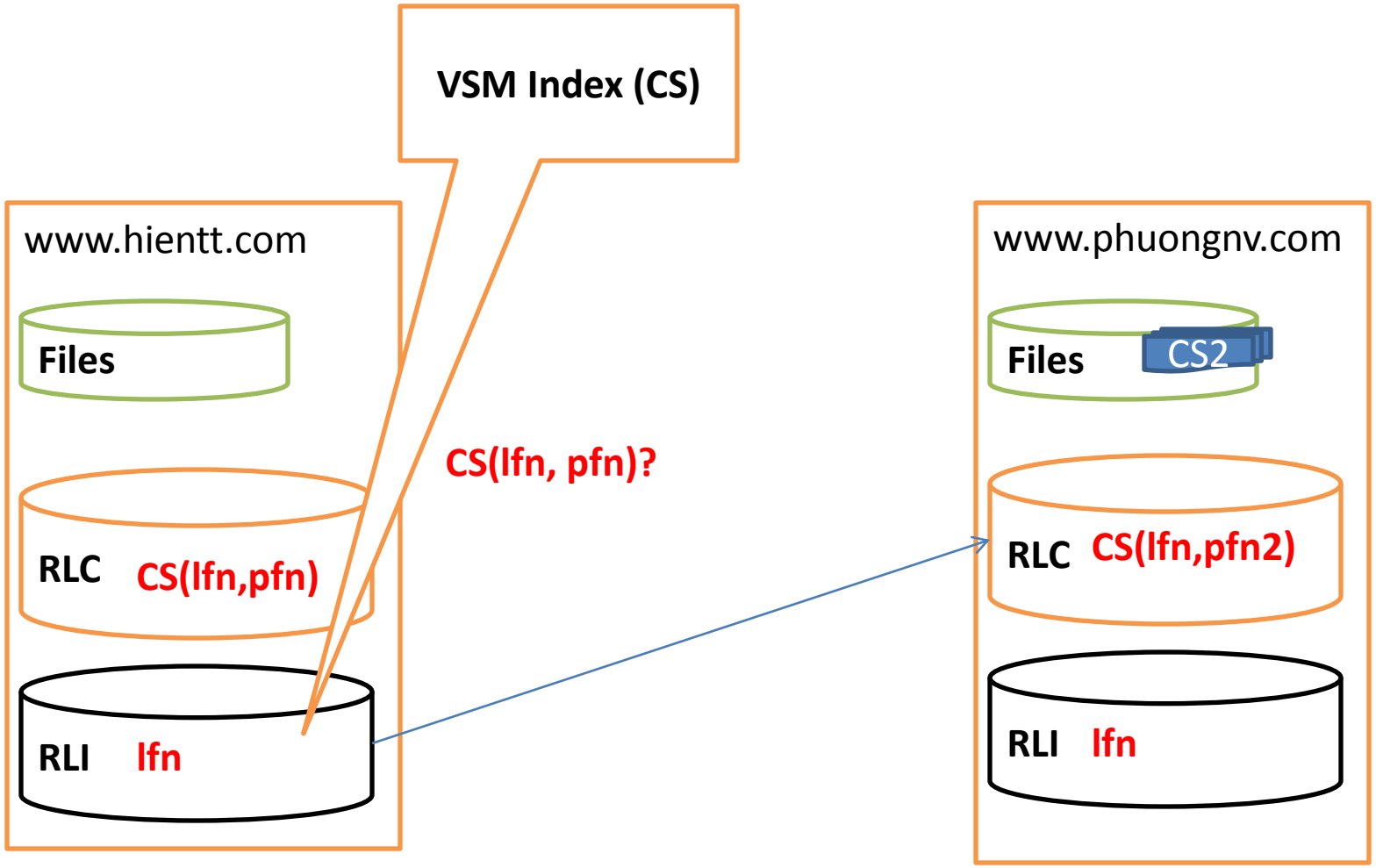
Kịch bản nhân bản dữ liệu

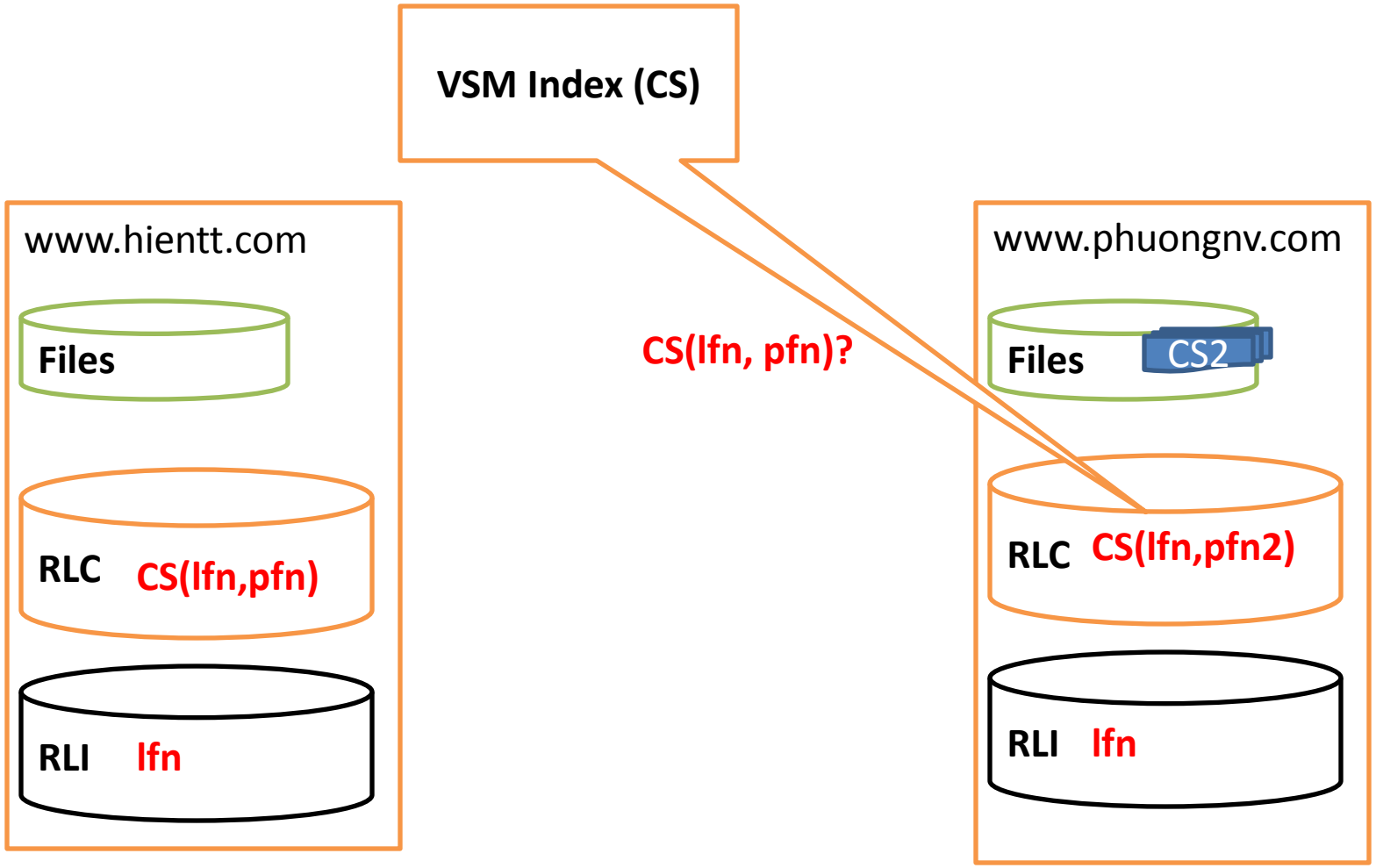
Những đóng góp của đồ án

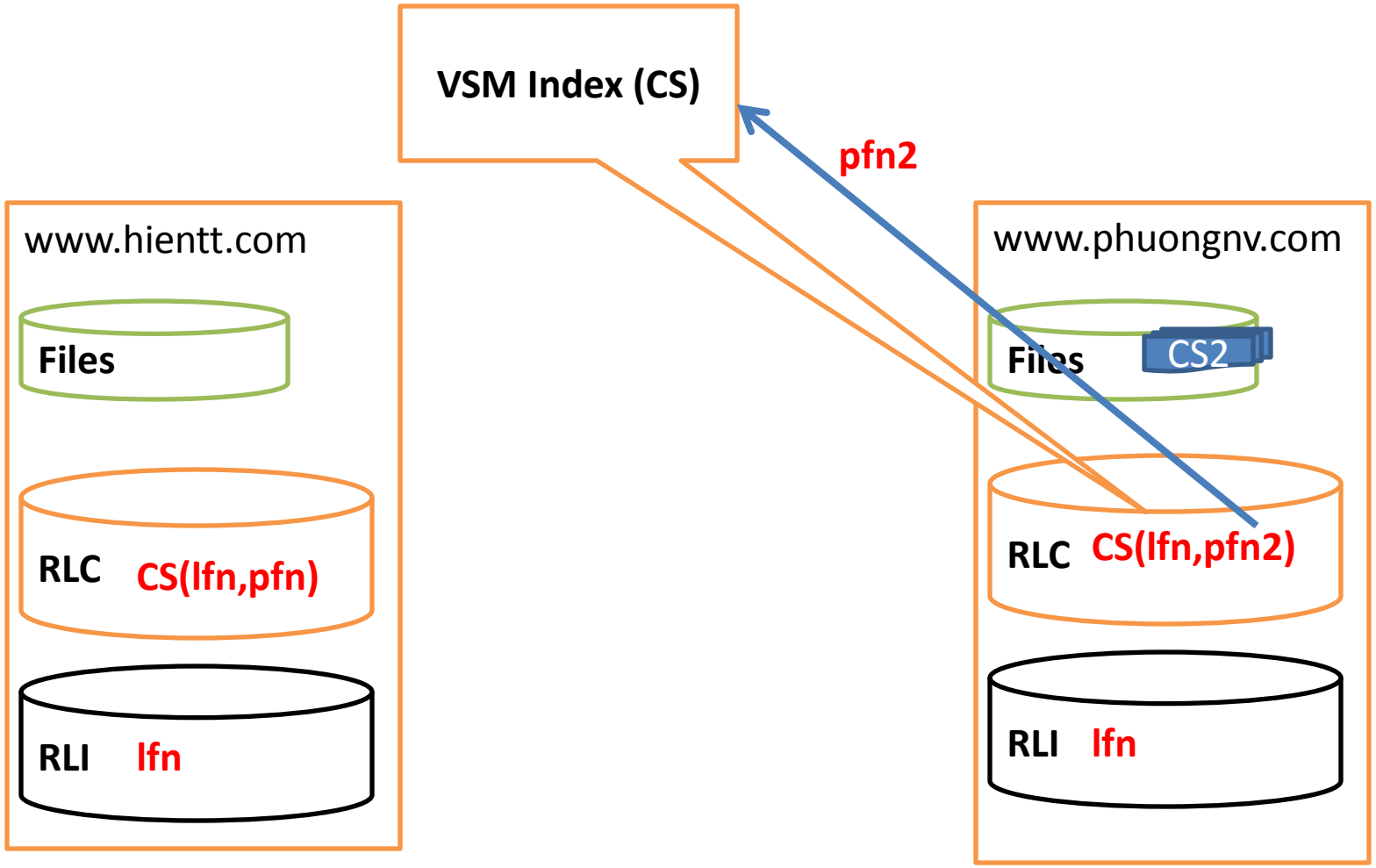
- Phân loại tài liệu để tăng hiệu quả của giải thuật PLSA




VSM Index (CS)
CS(lfn, pfn)?
www.hientt.com
Files
RLC
CS(lfn,pfn)
RLI
lfn
www.phuongnv.com
Files
CS2
RLC
CS(lfn,pfn2)
RLI
lfn


VSM Index (CS)
www.hientt.com
Files
RLC
CS(lfn,pfn)
RLI
lfn
CS(lfn, pfn)?
www.phuongnv.com
Files
CS2
RLC
CS(lfn,pfn2)
RLI
lfn


VSM Index (CS)
www.hientt.com
Files
RLC
CS(lfn,pfn)
RLI
lfn
CS(lfn, pfn)?
www.phuongnv.com
Files
CS2
RLC
CS(lfn,pfn2)
RLI
lfn


VSM Index (CS)
pfn2
www.hientt.com
Files
RLC
CS(lfn,pfn)
RLI
lfn
www.phuongnv.com
Files
CS2
RLC
CS(lfn,pfn2)
RLI
lfn