|  |  |
| --- | --- |
| Main Authors | C. Lee Giles, Kurt D. Bollacker, Steve Lawrence |
| Affiliations |  |
| Publisher | ACM Press |
| Conferences/Journal | Third ACM Conference on Digital Libraries |
| Public year | 1998 |
| Problem | * Cùng với sự phát triển của internet , số lượng các bài báo khoa học, số tạp chí điện tử, cũng như số các tác giả viết báo khoa học ngày càng tăng, điều này gây ra một số khó khăn khi mà người dùng muốn tìm bài báo về một vấn đề nào đó - cũng như tạo ra một thách thức lớn với các hệ thống đánh dấu – lưu trữ chỉ mục các bài báo trong việc cập nhật các bài báo một cách nhanh nhất. * Bài báo chỉ ra một số hệ thống đánh chỉ mục đã tồn tại và những ưu điểm cũng như giới hạn của nó:   + ISI (Institute Scientific information – Viện thông tin kkhoa học) : để xây dựng cơ sở dữ liệu chỉ mục, viện này sử dụng một số công cụ để đánh chỉ mục các bài báo trong các tạp chí khoa học, SCI (Science Citation Index) là một trong những công cụ mà viện sử dụng. Bài báo chỉ ra rằng database của ISI có một số nhược điểm, ví dụ database không bao quát hết các các hội nghị - các tạp chí , database không lưu trữ được các thông tin về các trích dẫn của các bài báo.    + Camenon : là một cơ sở dữ liệu chỉ mục các bài báo,trong đó hệ thống cho phép người dùng tìm kiếm thông tin các bài báo. Cơ sở dữ liệu này được thu thập và xây dựng bằng chính thông tin của tác giả bài báo, hoặc các tổ chức cung cấp. Như vậy nhược điểm của hệ thống là các giả phải cung cấp cho hệ thống thông tin về bài báo của mình theo đúng một định dạng (format) mà hệ thống đưa ra .   * Như vậy đối với những hệ thống đánh dấu – lưu trữ dữ liệu chỉ mục các bài báo khoa học hiện tại chưa đảm bảo được tính tự động và cập nhật dữ liệu. |
| Experimental Data | Dữ liệu kiểm thử được lấy từ 5093 tài liệu về “Neural Networks” và sau đây là kết quả từ hệ thống:   |  |  | | --- | --- | |  |  | | Số tài liệu  Số Citations nhận diện được  Số Titles nhận diện được  Số Authors nhận diện được  Số trang của bài báo nhận diện được (page) | 5093  89614  71908/89614 = 80.2%  73539/89614 = 82.1%  39595/89614 = 44.2% | |
| Main Contribution | CiteSeer là một hệ thống thu thập thông tin dữ liệu chỉ mục các bài báo khoa học được thu thập về từ trên web dựa trên nội dung và phần trích dẫn của bài báo, sau đó tổ chức thông tin dữ liệu chỉ mục thu thập được thành cơ sở dữ liệu, cho phép người dùng tìm kiếm thông tin trên dữ liệu chỉ mục này.  Sau đây là cách mà CiteSeer thực hiện:    CiteSeer sẽ download những bài báo có trên Internet, sau đó chuyển những bài báo theo các định dạng khác nhau về chung một định dạng text. Bước tiếp theo hệ thống sử dụng thuật toán và các mẫu (pattern) để rút trích thông tin chỉ mục của bài báo . Từ dữ liệu chỉ mục này xây dựng lên database của hệ thống.  + Hệ thống sử dung các web search enghines (Alta vista, Hotbot…) hoặc các thuật heuristic ( tìm kiếm với các từ khóa “papers”,”publications”… ) để tìm kiếm các bài báo trên web. Hệ thống sẽ download các Postscript hoặc file PDF của bài báo vừa tìm được về (hệ thống có cơ chế loại bỏ các bài báo đã download).  + Bài báo sau khi download về sẽ được chuyển định dang sang file Text. Bước tiếp theo hệ thống sẽ sử dụng một chương trình ( PreScript ) để xác định xem nội dung bài báo vừa download về có có phải là một tài liệu nghiên cứu không (ví dụ kiểm tra xem bài báo đó có resference không).  + Sau khi hệ thống nhận diện bài báo là một bài báo về nghiên cứu (chứ không phải là các bài báo trong lĩnh vực khác), hệ thống sẽ thực hiện việc nhận diện, rút trích các thành phần chính của bài báo bằng cách sử dụng các thuật toán heuristic , luật và máy học để rút các trường này của bài báo. Nội dung rút trích từ một tài liệu bao gồm:   * URL: Rút trích từ đường dẫn download tài liệu. * Header: phần tựa đề (title), thông tin tác giả (author) của bài báo. * Phần tóm tắt của bài báo nếu có (abstract). * Phần giới thiệu (introduction). * Phần tham khảo (citations): danh sách các tài liệu mà bài báo tham khảo. Cũng từ danh sáh này mà hệ thống sẽ phân tích và lưu trữ được thông tin bài báo này tham khảo những tài liệu này. * Nội dung bài báo.   Để giải quyết tình trạng 2 hoặc nhiều bài báo có phần trích dẫn đến cùng một tài liệu nhưng định dạng khác nhau (có thể điều gây lên khác biệt bao gồm: thứ tự tên tác giả trong bài báo khác nhau, hoặc 1 bên viết tắt các tên proceeding, conferece thành conf, proc .. một bên lại không viết tắt, hoặc có thể thêm các ký tự đặc biệt vào giữa các phần trong 1 reference), bài báo đưa ra các bước giải giải quyết như sau:  Từ phần citations rút trích được, đối với mỗi bài báo trích dẫn hệ thống sẽ:   * Chuyển font chữ của phần reference sang một loại font chữ. * Loại bỏ các đường dẫn liên kết nếu có trong mỗi mục. * Loại bỏ các tage đánh dấu thông tin refecence như : [1],[5] … * Mở rộng các phần tóm tắt nếu có trong phần reference ví dụ: conf ->conferece , proc. -> proceedings. * Loại bỏ các từ, các ký tự không mang ý nghĩa trong nội dung, ví dụ: pages, pp, …   Sau khi sử lý phần reference của bài báo hệ thống sẽ sử dụng thuật toán Likeit để so sánh và nhận diện xem bài phần bài báo trích dẫn này có liên quan đến bài báo nào có trong CSDL chưa.  Bài báo cũng giới thiệu chức năng tìm kiếm dựa trên dữ liệu chỉ mục thu thập được.  + Hệ thống cho phép tìm kiếm theo từ khóa (keyword) – hệ thống sẽ tìm kiếm những tài liệu có chưa từ khóa do người dùng nhập vào.  + Hệ thống cho phép tìm kiếm những tài liệu liên quan: hệ thống sử dụng các thông số như TFIDF và các thuật toán như Likeit để tìm kiếm và đưa ra kết quả cho người dùng. |
| Evaluation/ compare with related methods | Một số điểm chưa được:   * CiteSeer không thể đánh dấu dữ liệu chỉ mục với các tạp chí chưa có các bản điện tử. * CiteSeer chưa xử lý được những dữ liệu như sự nhập nhằng giữa 2 tác giả cùng một tên.   Một số điểm cải tiến so với các công cụ đã có:   * CiteSeer có thể đánh dấu dữ liệu chỉ mục các bài báo một cách nhanh nhất , ngay khi tác giả bài báo công bố bài báo lên trên web. * CiteSeer là một hệ thống tự động, hệ thống không cần thao tác trợ giúp thêm từ con người. |
| Expected Future works | * Cải tiến phần nhận diện các bài báo giống nhau. * Cải thiện các thuật toán nhận diện các trường trong thông tin dữ liệu chỉ mục. |

CiteSeer: An Automatic Citation Indexing System