MỤC LỤC

GIỚI THIỆU

Ngày nay, với sự phát triển mạnh mẽ của Internet, các hình thức kết nối và chia sẻ thông tin trong cộng đồng mạng ngày càng phát triển đã thu hút một lượng lớn người dùng tham gia. Qua đó, họ có thể dễ dàng trao đổi, chia sẻ thông tin, thảo luận các vấn đề và sở thích cùng quan tâm. Một số mạng xã hội phổ biến trên thế giới như: Facebook, Twitter, và ở Việt Nam như: Zing, Go.vn có số lượng người tham gia ngày càng đông đảo. Các bài nhận xét trên các diễn đàn, các trang dịch vụ và các trang tin tức cũng là một hình thức thể hiện khác rất phát triển.

Các thông tin được chia sẻ và thảo luận thông quan mạng xã hội thuộc rất nhiều chủ đề trong các lĩnh vực kinh tế, chính trị, xã hội. Từ đó hình thành nên các xu hướng, quan điểm của cộng đồng đối với việc đánh giá một vấn đề, hay một sản phẩm, dịch vụ nào đó. Các quan điểm, xu hướng này sẽ có tác động mạnh mẽ đến định hướng, quan điểm của người dùng khác

Trước đây, một người dùng khi muốn mua một sản phẩm hay sử dụng dịch vụ nào đó thường có xu hướng tìm hiểu thông tin qua những người xung quanh. Nhưng với sự phát triển của Internet như hiện nay, họ lại thường tìm các thông tin này qua mạng. Ví dụ như một người khi mua máy tính sẽ tìm hiểu thông tin về các sản phẩm trên mạng, thông thường sẽ chú ý đến các loại sản phẩm mà đa số người sử dụng đánh giá tốt, các loại sản phẩm được đề cập nhiều; một người đi du lịch sẽ chọn khách sạn có các tiêu chí quan tâm được cộng đồng đánh giá tích cực.

Mặt khác, đối với các nhà sản xuất, các nhà cung cấp dịch vụ để tìm hiểu các đánh giá của người dùng về sản phẩm và dịch vụ của mình, thay vì phải lấy phiếu điều tra cho sản phẩm một cách thủ công, họ có thể thu thập các thông tin thống kê quan điểm, xu hướng người dùng thông qua các trang mạng. Từ đó sẽ giúp các nhà sản xuất, các nhà cung cấp dịch vụ hoạch định các chính sách cần thiết để phát triển sản phẩm và đáp ứng phù hợp nhu cầu của thị trường.

Để có thể khai thác được các thông tin quan điểm của người dùng, việc tìm kiếm, trích các thông tin có liên quan đến các sản phẩm, dịch vụ có ý nghĩa quan trọng phục vụ cho hệ thống xử lý, đánh giá các quan điểm về sản phẩm, dịch vụ mà người dùng hay nhà sản xuất quan tâm.

Cũng vì lý do đó, trong đồ án này, em nghiên cứu về phương pháp trích và sắp xếp các đặc trưng của sản phẩm, từ đó có thể xác định các quan điểm hay nhận xét tới đặc trưng của sản phẩm đó, phục vụ cho các bước xử lý quan điểm tiếp theo.

Nội dung đồ án bao gồm 3 chương

Chương 1: Giới thiệu về bài toán phân tích quan điểm

Chương 2: Một số phương pháp trích và sắp xếp đặc trưng

Chương 3: Giới thiệu chương trình thực nghiệm và kết quả

Cuối cùng là phần kết luận

CHƯƠNG 1: BÀI TOÁN PHÂN TÍCH QUAN ĐIỂM

1.1 Nhu cầu về thông tin quan điểm và nhận xét

"Những gì người khác nghĩ" đã luôn luôn là một phần quan trọng trong việc cung cấp thông tin cho quá trình ra quyết định của hầu hết chúng ta. Trước khi Internet trở lên phổ biến, chúng ta thường yêu cầu bạn bè hay người thân giới thiệu một thợ cơ khí tự động hoặc yêu cầu tài liệu tham khảo liên quan đến xin việc từ các đồng nghiệp, hoặc tư vấn tiêu dùng. Ngày nay, Internet và Web đã giúp cho chúng ta có thể dễ dàng tiếp cận các ý kiến và kinh nghiệm của những người khác mà không nhất thiết phải là những người quen biết cá nhân, không phải là các nhà phê bình chuyên nghiệp nổi tiếng, những người mà chúng ta chưa bao giờ nghe nói tới trong không gian rộng lớn. Và ngược lại, ngày càng nhiều và nhiều hơn nữa những người sẵn sàng cung cấp các ý kiến của mình cho những người khác qua Internet.

Theo hai cuộc khảo sát của hơn 2000 người Mỹ trưởng thành mỗi: 81% người dùng Internet (hoặc 60% người Mỹ) đã thực hiện nghiên cứu trực tuyến về một sản phẩm ít nhất một lần; 20% (15% của tất cả các người Mỹ) làm như vậy trong một ngày. Trong số các độc giả đánh giá trực tuyến của nhà hàng, khách sạn, và các dịch vụ khác nhau (ví dụ như, các cơ quan du lịch hoặc bác sĩ), giữa 73% và 87% báo cáo đánh giá đã có một ảnh hưởng đáng kể mua hàng của họ; Người tiêu dùng sẵn sàng trả từ 20% đến 99% một mục được đánh giá 5 sao cao hơn so với một mục đánh giá 4 sao. 32% đã cung cấp một đánh giá về một sản phẩm, dịch vụ thông qua một hệ thống xếp hạng trực tuyến, trong đó có 18% của công dân trực tuyến cao cấp, có đăng một bình luận trực tuyến hoặc xem xét về một sản phẩm hay dịch vụ.

Thống kê nhanh chỉ ra rằng việc tiêu thụ hàng hóa và dịch vụ không phải là động cơ duy nhất khi người dùng tìm kiếm hoặc thể hiện ý kiến trực tuyến. Sự cần thiết của những thông tin chính trị cũng là một yếu tố quan trọng. Ví dụ, trong một cuộc khảo sát hơn 2500 người Mỹ trưởng thành, Rainie và Horrigan

nghiên cứu có 31% người Mỹ - trên 60 triệu người - 2006 người dùng Internet vận động tranh cử, là những người thu thập thông tin về cuộc bầu cử năm 2006 trực tuyến và trao đổi nhận xét thông qua email. Trong số này:

• 28% nói rằng nguyên nhân chính cho các hoạt động trực tuyến này để thu nhận được quan điểm từ bên trong cộng đồng của họ, và 34% cho biết một lý do chính là để nhận được quan điểm từ bên ngoài cộng đồng của họ.

• 27% đã xem đánh giá trực tuyến cho sự tán thành hoặc xếp hạng của các tổ chức bên ngoài.

• 28% cho biết rằng hầu hết các trang web mà họ sử dụng để chia sẻ quan điểm, nhưng 29% nói rằng phần lớn các trang web mà họ sử dụng thách thức quan điểm của họ, chỉ ra rằng nhiều người không chỉ đơn giản là tìm kiếm để xác nhận các quan điểm có trước của họ.

• 8% đăng bình luận trực tuyến bình luận chính trị riêng của họ.

Đối với người dùng tìm kiếm sự tin cậy trong những lời khuyên và tư vấn trực tuyến quan tâm đến việc xây dựng một hệ thống mới để xử lý trực tiếp các quan điểm trước tiên là phân loại chúng. Theo Horrigan thống kê rằng trong khi đa số người sử dụng internet của Mỹ cho rằng kinh nghiệm tích cực trong nghiên cứu sản phẩm trực tuyến, 58% cho rằng thông tin trực tuyến là thiếu, khó tìm, khó hiểu và hoặc quá nhiều. Vì vậy, nhu cầu có một hệ thống để hỗ trợ người tiêu dùng tìm kiếm thông tin là rất cần thiết.

Các nhà cung cấp sản phẩm ngày càng chú ý hơn đến sự quan tâm mà người dùng cá nhân thể hiện trong các ý kiến trực tuyến về sản phẩm và dịch vụ, và sự ảnh hưởng như xu thế sử dụng.

Với sự bùng nổ của nền tảng Web 2.0 như các blog, diễn đàn thảo luận, peer-to-peer mạng, và các loại khác nhau của các mạng xã hội. . .

• Thống kê của Facebook: có hơn 500 triệu người dùng ở trạng thái hoạt động (active) mỗi người có trung bình 130 bạn (friends), trao đổi qua lại trên 900 triệu đối tượng.

• Twitter (5/2011): có hơn 200 triệu người dùng. Một ngày có hơn 300 nghìn tài khoản mới, trung bình hơn 190 triệu tin nhắn, xử lý trung bình khoảng 1,6 tỷ câu hỏi

• Ở Việt Nam: các mạng xã hội zing.vn, go.vn … thu hút được đông đảo người dùng tham gia.

Một lượng đông đảo người dùng gia tăng chưa từng có và có quyền chia sẻ kinh nghiệm và ý kiến của riêng họ về bất kỳ sản phẩm hoặc dịch vụ, là tích cực hay tiêu cực. Khi các công ty lớn đang ngày càng nhận ra, những tiếng nói của người tiêu dùng có thể vận dụng rất lớn ảnh hưởng trong việc hình thành ý kiến của người tiêu dùng khác, cuối cùng, để trung thành với thương hiệu của họ, họ quyết định mua, và vận động cho chính thương hiệu của họ.. . Công ty có thể đáp ứng với những hiểu biết của người tiêu dùng mà họ tạo ra thông qua điều khiển phương tiện truyền thông xã hội và phân tích các thông điệp maketing của họ, định vị thương hiệu, phát triển sản phẩm và các hoạt động phù hợp khác.

Tuy nhiên, các nhà phân tích ngành công nghiệp lưu ý rằng việc tận dụng các phương tiện truyền thông mới cho mục đích theo dõi hình ảnh sản phẩm đòi hỏi cần phải có công nghệ mới.

Các nhà tiếp thị luôn luôn cần giám sát các phương tiện truyền thông cho thông tin liên quan đến thương hiệu của mình cho dù đó là đối với các hoạt động quan hệ công chúng, vi phạm gian lận, hoặc tình báo cạnh tranh. Nhưng phân mảnh các phương tiện truyền thông và thay đổi hành vi của người tiêu dùng đã loại trừ các phương pháp giám sát truyền thống. Technorati ước tính rằng 75.000 blog mới được tạo ra mỗi ngày, cùng với 1,2 triệu bài viết mỗi ngày, trong đó có nhiều ý kiến người tiêu dùng thảo luận về sản phẩm và dịch vụ.

Vì vậy, không chỉ có cá nhân, mà các công ty, các tổ chức đều quan tâm đến một hệ thống có khả năng tự động phân tích quan điểm của người tiêu dùng.

1.2 Lịch sử của phân tích quan điểm và khai thác quan điểm

Lĩnh vực phân tích quan điểm (sentiment analysis) hay khai thác quan điểm (opinion mining) gần đây đã thu hút được sự quan tâm rộng rãi của các nhà nghiên cứu. Năm 2001 bắt đầu đánh dấu sự lan rộng nhận thức về các vấn đề nghiên cứu và cơ hội nâng cao phân tích tình cảm và khai thác quan điểm.

Các nhân tố được nghiên cứu gồm:

• Sự gia tăng của các phương pháp học máy, xử lý ngôn ngữ tự nhiên và khôi phục thông tin.

• Sự sẵn có của các tập dữ liệu đào tạo cho các thuật toán học máy, sự phát triển của Internet, cụ thể là sự phát triển của tập hợp các trang Web thu thập các ý kiến và quan điểm.

• Thực hiện những thách thức trí tuệ, thương mại và các ứng dụng thông minh trong lĩnh vực này.

Thuật ngữ khai thác quan điểm (Dave et al. 2003) là các công cụ khai thác quan điểm sẽ xử lý một tập hợp các kết quả tìm kiếm cho một đối tượng nhất định, sinh ra một danh sách các thuộc tính sản phẩm (chất lượng, đặc trưng, vv) và các quan điểm tổng hợp về chúng (kém, bình thường, tốt).

“Phân tích quan điểm” là cụm từ song song của "khai thác quan điểm" ở những khía cạnh nhất định (Das và Chen Tong, 2001). “Phân tích quan điểm" và "khai thác quan điểm" biểu thị cùng một lĩnh vực nghiên cứu.

1.3 Nhiệm vụ của phân tích quan điểm

Phân tích quan điểm là những nghiên cứu nhằm phát hiện ra quan điểm hay xu hướng của người dùng dựa trên các kỹ thuật liên quan đến vấn đề xử lý ngôn ngữ tự nhiên.Có hai hướng tiếp cận chính cho bài toán này là: Phân lớp quan điểm (Sentiment Classification) và trích quan điểm (Sentiment Extraction)

• Phân lớp quan điểm: Là bài toán khai thác các kỹ thuật để phân lớp các văn bản theo định hướng quan điểm (tích cực, tiêu cực hay trung lập).

• Trích quan điểm: bao gồm 3 nhiệm vụ chính là:

1. Trích các đặc trưng đối tượng có nhận xét trong mỗi quan điểm.
2. Xác định có hay không các quan điểm trong các đặc trưng là positive, negative hay neutral (phụ thuộc vào định dạng của các quan điểm)
3. Nhóm các cụm từ cùng nghĩa đặc trưng.

CHƯƠNG 2: PHƯƠNG PHÁP TRÍCH VÀ SẮP XẾP ĐẶC TRƯNG SẢN PHẨM

2.1 Giới thiệu khai thác đặc trưng

Một nhiệm vụ để khai thác những quan điểm của con người biểu thị trên các đặc trưng của những thực thể. Làm thế nào để khai thác các đặc trưng từ một kho ngữ liệu là một vấn đề quan trọng. Đã có một số nghiên cứu về khai thác đặc trưng (Hu và Liu, 2004; Popescu và Etzioni, 2005; Kobayashi v.v..., 2007; Scaffidi v.v..., 2007; Stoyanov và Cardie, 2008; Wong v.v..., 2008; Qiu v.v..., 2009).

Phương pháp **Double propagation** (truyền kép) (Qiu v.v..., 2009) là một kỹ thuật không giám sát tiên tiến cho việc giải quyết vấn đề. Nó chủ yếu trích các đặc trưng là danh từ, hoạt động tốt trong ngữ liệu có kích thước trung bình. Tuy nhiên, phương pháp này có thể chỉ ra rất nhiều dữ liệu thừa (độ chính xác thấp), và nó có thể bỏ lỡ các đặc trưng quan trọng. Để đối phó với hai vấn đề này, Zhang và các cộng sự đề xuất một phương pháp khai thác đặc trưng mới. Họ cải tiến cho phương pháp của Qiu và các cộng sự., 2009 dựa trên mẫu bộ phận – toàn bộ và mẫu "No" được giới thiệu để tăng độ hồi tưởng. Bộ phận - toàn bộ hay meronymy là quan hệ ngữ nghĩa quan trọng trong NLP, mà chỉ ra rằng một hoặc nhiều đối tượng là một phần của một đối tượng khác.

Quan hệ này rất hữu ích cho khai thác đặc trưng, bởi vì nếu chúng ta biết một đối tượng là một phần của một lớp sản phẩm, đối tượng này cần phải là một đặc trưng. Mẫu "No" là một mẫu khai thác. Dạng cơ bản của nó là từ "No" đi theo sau bởi một danh từ/cụm danh từ. Mọi người thường biểu thị những bình luận ngắn hay những quan điểm của họ về các đặc trưng sử dụng mẫu này. Cả hai kiểu của những mẫu có thể giúp tìm thấy các đặc trưng bị mất bởi sự lan truyền. Đối với vấn đề độ chính xác thấp, họ giới thiệu một đặc trưng cấp phương pháp tiếp cận để giải quyết nó. Họ sắp xếp đặc trưng những ứng cử viên dựa vào sự quan trọng của họ bao gồm hai yếu tố: mức độ liên quan và tần suất

đặc trưng.

Ý tưởng cơ bản của sắp xếp tầm quan trọng đặc trưng là nếu một ứng cử viên đặc trưng là chính xác và thường xuyên được đề cập trong một kho ngữ liệu, nó cần phải được sắp xếp cao, nếu không nó phải là kết quả được sắp xếp thấp nhất trong kết quả cuối cùng. Tần suất đặc trưng là tần suất xuất hiện của một đặc trưng trong một kho ngữ liệu, mà dễ dàng để tồn tại. Tuy nhiên, đánh giá sự thích hợp đặc trưng là thách thức. Họ mẫu hóa các vấn đề như một đồ thị hai nhánh và sử dụng trang Web nổi tiếng sử dụng giải thuật HITS (Kleinberg, 1999) tới tìm kiếm tầm quan trọng đặc trưng và sắp xếp đặc trưng. Thử nghiệm của họ những kết quả cho thấy màn trình diễn vượt trội. Trong thực tế ứng dụng, họ tin rằng sắp xếp cũng quan trọng đối với khai thác đặc trưng vì sắp xếp có thể giúp người sử dụng khám phá các đặc trưng quan trọng từ hàng trăm kết quả những ứng cử viên có đặc trưng hiệu quả.

2.2 Một số phương pháp khai thác đặc trưng

Hu và Liu (2004) áp dụng việc gán nhãn từ loại và kĩ thuật xử lý ngôn ngữ tự nhiên nhằm rút ra những tính từ cũng như những từ chỉ quan điểm. Ý tưởng chính là những người thường sử dụng những từ ngữ giống nhau khi họ bình luận trên cùng những đặc trưng sản phẩm. Phương pháp của họ dựa vào việc phân loại dựa trên dấu hiệu quan điểm về sản phẩm:

- Định nghĩa một câu mà chứa một hay nhiều dấu hiệu sản phẩm và từ chỉ quan điểm được xem là một câu chỉ quan điểm.

- Với mỗi câu trong dữ liệu chỉ quan điểm, rút ra tất cả những tính từ được coi là những từ chỉ quan điểm.

- Kết quả thực nghiệm việc rút ra những câu đánh giá quan điểm có độ chính xác (precision) khoảng 64.2% và recall là 69.3%.

- Sử dụng WordNet (Fellbaum, 1998) để xác định các tính từ được rút ra mang chiều hướng tích cực (positive) hay tiêu cực (negative).

Trong WordNet, các tính từ được tổ chức thành các cụm từ lưỡng cực; nửa cụm thứ hai phần đầu là từ trái nghĩa của cụm thứ nhất. Mỗi nửa cụm là phần đầu của tập từ đồng nghĩa chính, tiếp theo là tập từ đồng nghĩa kèm theo – đại diện cho ngữ nghĩa tương tự như những tính từ quan trọng. Ngược với cách tiếp cận dựa trên từ điển, họ sử dụng định hướng quan điểm của những từ đồng nghĩa và từ trái nghĩa để dự đoán định hướng của các tính từ. Họ bắt đầu với một danh sách khởi đầu gồm 30 tính từ thông dụng được chọn thủ công (bằng tay). Sau đó sử dụng WordNet để dự đoán định hướng của tất cả các tính từ trong danh sách từ quan điểm được rút ra bằng cách tìm kiếm qua cụm lưỡng cực để tìm ra liệu các từ đồng nghĩa hay trái nghĩa có trong danh sách khởi đầu hay không. Khi định hướng của tính từ được dự đoán, nó sẽ được bổ sung vào danh sách khởi đầu và có thể được sử dụng để xác định định hướng của các tính từ khác. Trong phương pháp này, danh sách khởi đầu sẽ dần tăng lên khi sự định hướng của các tính từ được nhận dạng; và khi nó ngừng gia tăng, tức qui mô của danh sách khởi đầu trùng với qui mô của danh sách từ chỉ quan điểm, thì tất cả định hướng của các tính từ đã được nhận biết và quá trình này kết thúc.

Popescu và Etzioni (2005) nghiên cứu cùng một vấn đề. Giải thuật của họ yêu cầu lớp sản phẩm được biết đến. Giải thuật xác định liệu có phải một danh từ/ cụm danh từ là một đặc trưng bằng cách tính toán thông tin theo từng điểm tổng quan lẫn nhau (PMI) đánh dấu giữa mệnh đề và lớp. Đầu tiên sử dụng mẫu bộ phận - toàn bộ để khai thác đặc trưng, toàn bộ dựa trên các đặc trưng bằng cách tìm kiếm trên web. Truy vấn trên web cần nhiều thời gian. Trong phương pháp của họ, họ sử dụng những mẫu quan hệ bộ phận - toàn bộ đặt sẵn để trích các đặc trưng trong một miền ngữ liệu. Những mẫu này là miền độc lập và khá chính xác.

Sau nghiên cứu ban đầu (Hu và Liu, 2004), một số nhà nghiên cứu đã tiếp tục khám phá ý tưởng sử dụng những quan điểm trong khai thác đặc trưng sản phẩm. Một phương pháp được đề xuất dựa trên phần phụ thuộc (Zhuang, 2006) xem xét ứng dụng phân tích tổng quan. Qiu (2009) đề xuất phương pháp double

propagation, khai thác các mối quan hệ cú pháp nhất định của phát biểu quan điểm và làm nối bật, lan truyền thông qua cả những từ quan điểm lẫn các đặc trưng lặp đi lặp lại. Những quy tắc khai thác được thiết kế tiếp tục đặt cơ sở những quan hệ khác nhau giữa các quan điểm và các đặc trưng. Ngữ pháp phụ thuộc đã được thông qua mô tả những mối quan hệ này. Với Wang (2008) một phương pháp nạp chương trình nguồn được đề xuất. Với Kobayashi (2007) một mẫu phương pháp khai thác được sử dụng. Các mẫu là những quan hệ giữa đặc trưng và những cặp quan điểm. Những mẫu được khai thác từ một kho ngữ liệu lớn bằng cách sử dụng mẫu khai thác mẫu. Thống kê từ kho ngữ liệu được sử dụng để xác định những điểm tin cậy của việc khai thác.

Khai thác thông tin tổng quát có hai cách tiếp cận: dựa trên nguyên tắc và thống kê. Những hệ thống khai thác đầu tiên chủ yếu dựa trên các quy tắc (Riloff, 1993). Trong những phương pháp thống kê, hầu hết các mẫu phổ biến là Hidden Markov Models (HMM_Rabiner, 1989), Maximum Entropy Models (MEM_Chieu, 2002) và Conditional Random Fields (CRF_Lafferty, 2001).

2.3 Phương pháp trích và sắp xếp các đặc trưng quan điểm về sản phẩm.

Phương pháp này giả thiết rằng các đặc trưng là danh từ/ cụm danh từ và các từ quan điểm là các tính từ. Điều này cho thấy các từ quan điểm thường liên kết với các đặc trưng theo một số cách nào đó. Do đó,những từ quan điểm có thể được nhận biết qua các đặc trưng đã xác định, và các đặc trưng có thể được xác định những từ quan điểm đã biết. Các quan điểm và các đặc trưng đã được trích được sử dụng để xác định những quan điểm và đặc trưng mới, rồi chúng lại được sử dụng một lần nữa để khai thác những quan điểm và các đặc trưng nhiều hơn. Sự lan truyền hay quá trình bootstrapping kết thúc khi không có các từ quan điểm hay đặc trưng nào có thể được tìm thấy. Ưu điểm lớn nhất của phương pháp này là nó không đòi hỏi nguồn tài nguyên bổ sung ngoại trừ một từ qua điểm giống được khởi tạo ban đầu. Vì vậy, phương pháp này là độc lập với miền dữ liệu và là phương pháp không giám sát, tránh mất thời gian và gán nhãn dữ liệu như các phương pháp học không giám sát. Nhưng với ngữ liệu lớn,

phương pháp này có thể trích nhiều danh từ/ cụm danh từ không phải là đặc trưng. Độ chính xác của phương pháp do đó giảm xuống. Lý do là trong quá trình truyền, những tính từ không là quan điểm vẫn được trích như là quan điểm, ví dụ như *"entire"* và *"current"*. Các tính từ này không là quan điểm, nhưng chúng có thể bổ nghĩa cho một số loại danh từ, cụm danh từ, do đó dẫn tới trích các đặc trưng sai. Lặp đi lặp lại, càng ngày càng nhiều những dữ liệu nhiễu có thể được thực hiện trong suốt quá trình. Các vấn đề khác là cho những miền dữ liệu nhất định, một số đặc trưng quan trọng không có các từ quan điểm bổ sung cho chúng.

Ví dụ: trong một nhận xét về các sản phẩm đệm: *"There is a valley on my mattress"(có một cái vũng trên đệm của tôi)*, ngụ ý một quan điểm tiêu cực vì *"valley" (vũng)* là điều không mong muốn trên *"mattress"(đệm)*. Rõ ràng, *"valley"* là một đặc trưng nhưng nó không được mô tả bởi một tính từ quan điểm, đặc biệt là cho một ngữ liệu nhỏ. Phương pháp Double propagation không thích hợp cho trường hợp này.

Để giải quyết vấn đề này, Zhang và các cộng sự đã đề xuất phương pháp mới để khai thác đặc trưng bao gồm hai bước sau: khai thác đặc trưng và sắp xếp đặc trưng. Để khai thác đặc trưng họ vẫn áp dụng ý tưởng double propagation để xác định những ứng cử viên. Nhưng có hai cải tiến dựa trên quan hệ bộ phận - toàn bộ (part-whole relation) và mẫu "No"("No" pattern) được thực hiện tìm kiếm các đặc trưng mà double propagation không thể tìm thấy. Chúng có thể giải quyết một phần vấn đề độ hồi tưởng (recall). Đối với sắp xếp đặc trưng, các tác giả đã sắp xếp các ứng cử viên đặc trưng dựa trên tầm quan trọng của chúng.

Một mẫu bộ phận - toàn bộ cho thấy một đối tượng là một phần của một đối tượng khác. Ở ví dụ trên: *"There is a valley on my mattress"*, chúng ta có thể thấy nó bao gồm quan hệ bộ phận - toàn bộ giữa *"valley"* và *"mattress"*, *"valley"* quan hệ với *"mattress"*, nó được chỉ ra bởi giới từ *"on"*. *"Valley"* không thực sự là một phần của *"mattress"*, nhưng là một tác động trên mattress.

Nó được gọi là mối quan hệ bộ phận - toàn bộ giả (a pseudo part-whole relation). Để đơn giản, chúng ta không phân biệt nó với mối quan hệ bộ phận - toàn bộ thực tế bởi vì với nhiệm vụ khai thác đặc trưng, chúng khác biệt rất ít. Trong trường hợp này, "noun1 on noun2" là một mẫu tốt, nó ngụ ý noun1 là một phần của noun2. Nếu chúng ta biết *"mattress"* là một khái niệm lớp, chúng ta có thể suy ra rằng *"valley"* là một đặc trưng cho *"mattress"*. Có rất nhiều các cụm từ hoặc các mẫu câu thể hiện dạng này của mối quan hệ ngữ nghĩa đã được nghiên cứu (Girju et al., 2006). Bên cạnh các mẫu quan hệ bộ phận - toàn bộ, mẫu "No" là một mẫu quan trọng khác và chỉ ra các đặc trưng đặc biệt trong tài liệu chứa quan điểm.

Để giải quyết vấn đề đầu tiên: dữ liệu nhiễu, với những quan điểm, mẫu bộ phận - toàn bộ và mẫu "No", các tác giả có ba chỉ số đặc trưng trong tay, nhưng tất cả đều là không rõ ràng, có nghĩa rằng chúng không phải là các luật khó. Chúng ta sẽ không tránh khỏi khai thác các đặc trưng sai (còn gọi là dữ liệu nhiễu) bằng cách sử dụng chúng. Cắt bỏ những dữ liệu nhiễu từ những ứng cử viên đặc trưng là một nhiệm vụ khó khăn. Thay vào đó, các tác giả đề xuất một cách để giải quyết vấn đề này: sắp xếp đặc trưng.

Ý tưởng cơ bản là chúng ta sắp xếp những ứng cử viên đặc trưng được trích bởi tầm quan trọng đặc trưng. Nếu một ứng cử viên là đặc trưng chính xác và quan trọng, nó phải được sắp xếp cao. Đối với đặc trưng không quan trọng hoặc nhiễu, nó phải được sắp xếp với hạng thấp trong kết quả cuối cùng. Bảng sắp xếp cũng rất hữu ích trong thực tế. Trong một kho ngữ liệu lớn, chúng ta có thể rút ra hàng trăm đặc trưng nổi bật. Tuy nhiên, người sử dụng thường chỉ quan tâm về những vấn đề quan trọng, chúng cần phải có thứ hạng cao. Các tác giả xác định hai nhân tố chính ảnh hưởng đến tầm quan trọng đặc trưng: một là sự thích hợp đặc trưng (feature relevance) và hai là tần suất đặc trưng (feature frequency).

Thích hợp đặc trưng: nó mô tả làm thế nào có thể xảy ra một ứng cử viên đặc trưng là một đặc trưng chính xác. Các tác giả thấy rằngcó ba đầu mối mạnh mẽ chỉ sự thích hợp đặc trưng trong một kho ngữ liệu.

- ✓ Đầu mối đầu tiên là một đặc trưng chính xác thường được bổ nghĩa bởi nhiều từ quan điểm (các tính từ hay các trạng từ). Ví dụ, trong dữ liệu về mattress (đệm), *"delivery"* được bổ nghĩa bởi *"quick" "cumbersome"* và *"timely"*. Nó cho thấy nhận xét này nhấn mạnh vào từ *"delivery"*. Do đó chúng ta có thể suy luận rằng *"delivery"* là một đặc trưng phù hợp.
- ✓ Đầu mối thứ hai là một đặc trưng có thể được rút ra từ nhiều các mẫu bộ phận - toàn bộ. Ví dụ, trong dữ liệu ô tô, nếu chúng ta tìm thấy hai cụm từ sau, *"the engine of the car"* và *"the car has a big engine"*, chúng ta có thể suy luận rằng *"engine"* là một phần của *"car"*.
- ✓ Đầu mối thứ ba là sự kết hợp của sự bổ nghĩa quan điểm, trích mẫu bộ phận - toàn bộ và mẫu "No". Đó là, nếu một ứng cử viên đặc trưng không chỉ được bổ nghĩa bởi các từ quan điểm mà còn được trích ra từ mẫu bộ phận - toàn bộ hay mẫu "No", chúng ta có thể suy luận đó là một đặc trưng với độ tin cậy cao. Ví dụ, câu *"there is a bad hole in the mattress"*, nó chỉ ra một cách rõ ràng là *"hole"* là một đặc trưng cho mattress vì nó được bổ nghĩa bởi từ quan điểm *"bad"* và cũng trong mẫu bộ phận - toàn bộ.

Ngoài ra, các tác giả thấy rằng có một mối quan hệ thực thi lẫn nhau giữa các từ quan điểm, các mẫu bộ phận - toàn bộ và "No", và các đặc trưng. Nếu một tính từ bổ nghĩa cho nhiều đặc trưng đúng, thì rất có thể là từ những quan điểm tốt. Tương tự, nếu một ứng cử viên đặc trưng có thể được rút ra dựa trên nhiều từ quan điểm, các mẫu bộ phận - toàn bộ, hoặc mẫu "No", nó cũng có khả năng cao là một đặc trưng đúng. Điều này chỉ ra rằng giải thuật HITS sắp xếp các trang Web có thể áp dụng được.

Tần suất đặc trưng: đây là một yếu tố quan trọng ảnh hưởng đến việc sắp xếp đặc trưng. Tần xuất đặc trưng đã được xem xét trong nghiên cứu của Hu và Liu, năm 2004; Blair-Goldensohn và các cộng sự năm 2008. Các tác giả cho rằng một đặc trưng f1 thì quan trọng hơn đặc trưng f2 nếu f1 xuất hiện thường xuyên hơn so với f2 trong những tài liệu quan điểm. Trong thực tế, đó là mong muốn để sắp xếp các đặc trưng thường xuyên đó cao hơn so với các đặc trưng hiếm khi xảy ra. Nguyên nhân là thiếu một đặc trưng được đề cập thường xuyên trong khai thác quan điểm là xấu, nhưng thiếu một tỉ lệ đặc trưng không phải là một vấn đề lớn.

Kết hợp các nhân tố trên, các tác giả giới thiệu một phương pháp khai thác đặc trưng mới. Thực nghiệm cho kết quả tốt với nhiều tập dữ liệu thực tế đa dạng.

2.3.1 Double propagation

Double propagation dựa vào quan sát mà ở đó là quan hệ tự nhiên giữa các từ quan điểm và các đặc trưng vì thực tế là các từ quan điểm thường được sử dụng để bổ nghĩa cho các đặc trưng. Hơn nữa, quan sát cũng cho thấy rằng các từ quan điểm và các đặc trưng cuả chúng cũng có quan hệ trong các thể hiện chứa quan điểm (Qiu và các cộng sự năm 2009). Các mối quan hệ này có thể được xác định thông qua bộ phân tích cú pháp phụ thuộc dựa vào ngữ pháp phụ thuộc. Việc xác định các quan hệ này là chìa khoá để khai thác đặc trưng.

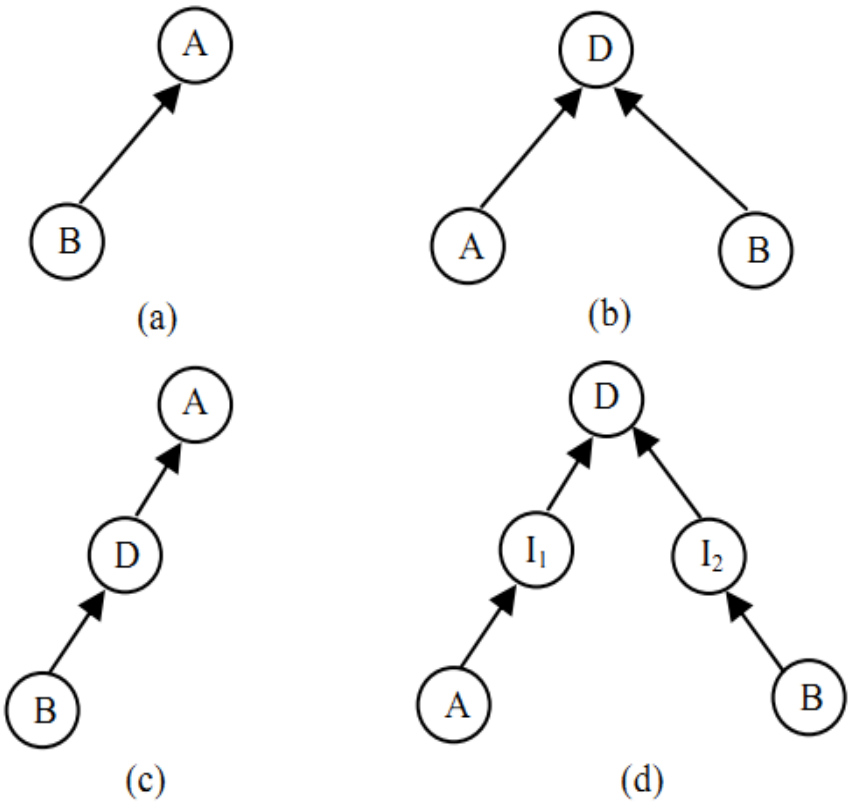
Dependency grammar (Ngữ pháp phụ thuộc): mô tả các quan hệ phụ thuộc giữa các từ trong một câu. Sau khi được phân tích bởi một phân tích cú pháp phụ thuộc, các từ trong một câu được liên kết với nhau bằng một quan hệ chắc chắn. Với câu, *"The camera has a good lens"*, *"good"* là từ quan điểm và *"lens"* là một đặc trưng của *camera*. Sau khi phân tích cú pháp, chúng ta có thể thấy rằng *"good"* phụ thuộc vào *"lens"* với mối quan hệ mod. **Mod** có nghĩa là *"good"* là từ bổ nghĩa cho *"lens"*. Trong một số trường hợp, một từ quan điểm và đặc trưng không trực tiếp phụ thuộc, nhưng chúng phụ thuộc trực tiếp vào cùng một từ. Ví dụ, từ câu *"The lens is nice"* chúng ta có thể tìm thấy cả đặc

trưng *"lens"* và từ quan điểm *"nice"* phụ thuộc vào động từ *"is"* với mối quan hệ **s** và pred tương ứng. Ở đây **s** có nghĩa là *"lens"* là đối tượng bề mặt của *"is"* trong khi **D** có nghĩa *"nice"* là vị ngữ của mệnh đề.

Trong (Qiu và các cộng sự., 2009), định nghĩa hai phạm trù quan hệ phụ thuộc để tổng kết tất cả các kiểu quan hệ phụ thuộc giữa hai từ, được minh họa trong hình 1. Mũi tên được sử dụng để đại diện cho những phần phụ thuộc.

Quan hệ trực tiếp (Direct relations - DR): Nó đại diện cho một từ phụ thuộc vào từ khác trực tiếp hoặc cả hai đều phụ thuộc trực tiếp vào một từ thứ ba, thể hiện trong (a) và (b) của hình 1. Trong (a), B phụ thuộc trực tiếp vào A, và trong (b) cả hai đều trực tiếp phụ thuộc vào D.

Quan hệ gián tiếp (Indirect relation): được thể hiện cho việc một từ phụ thuộc vào từ khác thông qua những từ khác hay cả hai phụ thuộc vào một từ thứ ba gián tiếp. Ví dụ, trong (c) của hình 1, B phụ thuộc vào A thông qua D; trong (d) hình 1, A phụ thuộc vào D thông qua I_1 trong khi B phụ thuộc vào D thông qua I_2. Đối với một số tình huống phức tạp, có thể có nhiều hơn một I_1,I_2.



Hình 1: Mối quan hệ khác nhau giữa A và B

Các mối quan hệ gián tiếp cú pháp là nghiêng về lỗi trong ngữ liệu Web. Do đó các tác giả chỉ sử dụng mối quan hệ trực tiếp để trích các từ quan điểm và các ứng cử viên đặc trưng trong ứng dụng. Sử dụng các luật trong Qiu và các cộng sự., 2009 được áp dụng như sau:

• Các luật trích dựa trên các mối quan hệ **(Extraction Rules based on Relations)**

Cho hai quan hệ trực tiếp DRs giữa A và B (cả A và B có thể là các từ quan điểm hoặc đặc trưng), chúng ta định nghĩa các luật để thu được các mối quan hệ cụ thể cũng như các thông tin từ loại. Sử dụng công cụ gán nhãn từ loại Stanford POS tagger. Với 4 loại của nhiệm vụ trích, chúng ta định nghĩa 4 luật như bảng 1.

Cột 2 là các mối quan hệ được quan sát giữa hai từ, Cột 3 thể hiện sự ràng buộc của mối quan hệ quan sát được và cột 4 là kết quả. Mũi tên thể hiện sự phụ thuộc. Ví dụ, S→S-Dep→F có nghĩa S phụ thuộc vào F thông qua một mối quan hệ S-Dep.

	Observations	Constraints	Outputs
$R1_1$	$S_{i(j)} \rightarrow S_{i(j)}\text{-}Dep \rightarrow S_{j(i)}$	$S_{j(i)} \in \{S\}$, $S_{i(j)}\text{-}Dep \in \{CONJ\}$, $POS(S_{i(j)}) \in \{JJ\}$	$s = S_{i(j)}$
$R1_2$	$S_i \rightarrow S_i\text{-}Dep \rightarrow H \leftarrow S_j\text{-}Dep \leftarrow S_j$	$S_i \in \{S\}$, $S_i\text{-}Dep == S_j\text{-}Dep$, $POS(S_j) \in \{JJ\}$	$s = S_j$
$R2_1$	$S \rightarrow S\text{-}Dep \rightarrow F$	$F \in \{F\}$, $S\text{-}Dep \in \{MR\}$, $POS(S) \in \{JJ\}$	$s = S$
$R2_2$	$S \rightarrow S\text{-}Dep \rightarrow H \leftarrow F\text{-}Dep \leftarrow F$	$F \in \{F\}$, $S/F\text{-}Dep \in \{MR\}$, $POS(S) \in \{JJ\}$	$s = S$
$R3_1$	$S \rightarrow S\text{-}Dep \rightarrow F$	$S \in \{S\}$, $S\text{-}Dep \in \{MR\}$, $POS(F) \in \{NN\}$	$f = F$
$R3_2$	$S \rightarrow S\text{-}Dep \rightarrow H \leftarrow F\text{-}Dep \leftarrow F$	$S \in \{S\}$, $S/F\text{-}Dep \in \{MR\}$, $POS(F) \in \{NN\}$	$f = F$
$R4_1$	$F_{i(j)} \rightarrow F_{i(j)}\text{-}Dep \rightarrow F_{j(i)}$	$F_{j(i)} \in \{F\}$, $F_{i(j)}\text{-}Dep \in \{CONJ\}$, $POS(F_{i(j)}) \in \{NN\}$	$f = F_{i(j)}$
$R4_2$	$F_i \rightarrow F_i\text{-}Dep \rightarrow H \leftarrow F_j\text{-}Dep \leftarrow F_j$	$F_i \in \{F\}$, $F_i\text{-}Dep == F_j\text{-}Dep$, $POS(F_j) \in \{NN\}$	$f = F_j$

Bảng 1: Các luật để trích từ quan điểm và đặc trưng.

Trong bảng, s(or f) nghĩa là từ quan điểm được trích (hoặc đặc trưng). {S}(or {F}) và S(or F)-Dep viết tắt cho các từ quan điểm đã biết (hoặc các đặc trưng đã được trích) và mối quan hệ phụ thuộc của S(or F) theo thứ tự tương ứng. H có nghĩa là một từ bất kỳ. POS(S(or F)) là thông tin từ loại của S(or F). {JJ} và {NN} là tập các nhãn từ loại của các từ quan điểm và đặc trưng tiềm năng tương ứng (JJ: là nhãn từ loại tính từ và NN: là nhãn từ loại danh từ). Các tác giả xem xét các từ quan điểm là các tính từ như trong phần lớn các nghiên cứu trước về phân tích quan điểm. và các đặc trưng là danh từ /cụm danh từ. Do đó, {JJ} ban gồm JJ, JJR (các tính từ dạng so sánh hơn) và JJS (các từ dạng so

sánh hơn nhất). {NN} bao gồm NN và NNS, là viết tắt cho danh từ số ít và danh từ số nhiều. Tuy nhiên, có các trường hợp mà các nhận xét sử dụng các đại từ để tham chiếu đến các đặc trưng đã được đề cập trước đó. Do đó, các tác giả cũng xem xét các đại từ như là các đặc trưng. Trong đó, họ sử dụng *"it"* và *"they"*. Do các lỗi là có khả năng, các tác giả không thực hiện bất kỳ xử lý tham chiếu đồng thời nào.

{MR} bao gồm các mối quan hệ phụ thuộc mô tả các mối quan hệ giữa các từ quan điểm và các đặc trưng, như là mod, có nghĩa là một từ bổ nghĩa cho một từ khác. Các mối quan hệ phụ thuộc khác (MRs) bao gồm: subj, obj, pnmod, etc.

{CONJ} là mối quan hệ của liên từ và chỉ bao gồm liên từ.

Trong đó sử dụng:

- ✓ $R1_i$ để trích các từ quan điểm (s) sử dụng các từ quan điểm words (S_i)
- ✓ $R2_i$ để trích các từ quan điểm (s) sử dụng các đặc trưng (F)
- ✓ $R3_i$ để trích các đặc trưng (f) sử dụng các từ quan điểm (S)
- ✓ $R4_i$ để trích các đặc trưng (f) sử dụng các đặc trưng đã được trích (F_i).

2.3.2 Mối quan hệ bộ phận - toàn bộ (Part-whole relation)

Một mối quan hệ bộ phận - toàn bộ là một tiêu chí tốt cho các đặc trưng nếu từ khái niệm lớp được biết đến. Ví dụ, một kết hợp định danh *"car hood"* bao gồm mối quan hệ bộ phận - toàn bộ. Nếu chúng ta biết *"car"*là một từ khái niệm lớp được biết đến, thì ta có thể suy luận rằng *"hood"* là một đặc trưng của *car*. Mẫu toàn bộ xuất hiện thường xuyên trong văn bản và được thể hiện bởi sự đa dạng của các cấu trúc cú pháp-từ vựng (lexico-syntactic structures) (Girju v.v..., 2006; Popescu và Etzioni, 2005). Có hai kiểu cấu trúc cú pháp-từ vựng truyền đạt các mối quan hệ bộ phận - toàn bộ: cấu trúc rõ ràng (unambiguous structure) và cấu trúc nhập nhằng (ambiguous structure). Cấu trúc rõ ràng chỉ ra rõ ràng mối quan hệ bộ phận - toàn bộ - từng phần. Ví dụ, câu *"the camera*

consists of lens, body and power cord." và *"the bed was made of wood"*. Trong những trường hợp này, sự phát hiện của các mẫu dẫn tới khám phá các mối quan hệ bộ phận - toàn bộ thực tế. Chúng ta có thể dễ dàng tìm thấy các đặc trưng của *camera* và *the bed*. Thật không may, điều này không phải loại mẫu thường xuyên trong ngữ liệu.

Tuy nhiên, có rất nhiều sự nhập nhằng thể hiện rõ ràng nhưng chuyển tải các mối quan hệ bộ phận - toàn bộ trong một số bối cảnh. Ví dụ, cho hai cụm từ *"valley on the mattress"* và *"toy on the mattress"*, *"valley"* là một phần của *"mattress"* trong khi *"toy"* không phải là một phần của *"mattress"*. Các tác giả sử dụng cả hai mẫu rõ ràng và nhập nhằng. Mặc dù các mẫu nhập nhằng có thể mang lại một số dữ liệu thừa, chúng ta có thể sắp xếp chúng với chỉ số thấp trong thủ tục sắp xếp. Hai loại mẫu sau được sử dụng để trích đặc trưng.

2.3.2.1 Mẫu cụm từ (Phrases pattern)

Trong trường hợp này, mối quan hệ bộ phận - toàn bộ tồn tại trong một cụm từ.

NP + Prep + **CP**: danh từ / cụm từ danh từ (NP) chứa đựng từ bộ phận và cụm khái niệm lớp (CP) có chứa từ toàn bộ. Chúng được nối bởi giới từ (Prep). Ví dụ, *"battery of the camera"* là một trường hợp của mẫu này NP (*battery*) là danh từ bộ phận và CP (*camera*) là danh từ toàn bộ. Trong ứng dụng của các tác giả, họ chỉ sử dụng 3 vị trí đặc biệt: *"of"*, *"in"* và *"on"*.

CP + with + **NP**: tương tự như vậy, CP là cụm khái niệm lớp, và NP là cụm danh từ / danh từ. Chúng được nối với nhau bằng từ *"with"*. Ở đây NP là có khả năng là một đặc trưng. Ví dụ, trong một mệnh đề, *"mattress with a cover"*, *"cover"* là một đặc trưng cho *mattress*.

NP CP hay **CP NP**: danh từ / cụm danh từ (NP) và cụm khái niệm lớp (CP) tạo thành một từ ghép. Ví dụ, *"mattress pad"*. *"Pad"* là một đặc trưng của *"mattress"*.

2.3.2.2 Mẫu câu (Sentence pattern)

Trong các mẫu này, quan hệ bộ phận – toàn bộ được chỉ định trong một câu. Các mẫu có chứa động từ cụ thể. Từ bộ phận có thể được tìm thấy bên trong các cụm danh từ hoặc các cụm giới từ chứa giới từ xác định.

"**CP** Verb **NP**": CP là cụm khái niệm lớp có chứa từ toàn bộ, NP là cụm danh từ có chứa các từ bộ phận và động từ có giới hạn và xác định. Ví dụ, trong một câu, *"the phone has a big screen"*, chúng ta có thể suy luận rằng *"screen"* là một đặc trưng cho *"phone"*, mà là một khái niệm lớp. Trong những mẫu câu, động từ đóng một vai trò quan trọng. Các tác giả sử dụng động từ thể hiện để tìm các quan hệ bộ phận - toàn bộ trong một câu, thí dụ, *"has", "have" "include" "contain" "consist", "comprise"*. .... (Girju v.v..., 2006).

Đây là đề cập hữu ích để sử dụng các mối quan hệ bộ phận – toàn bộ, từ khái niệm lớp cho ngữ liệu là cần thiết, nó khá dễ dàng tìm kiếm được bởi vì danh từ với tần xuất xuất hiện lớn trong ngữ liệu thì luôn luôn là từ khái niệm lớp dựa trên các thực nghiệm của các tác giả.

2.3.3 Mẫu "No"

Bên cạnh từ quan điểm và mối quan hệ bộ phận – toàn bộ, mẫu "No" cũng là một mẫu quan trọng cho thấy các đặc trưng trong một kho ngữ liệu. Ở đây "No" thể hiện cho từ no. Dạng cơ bản của mẫu "No" là từ theo sau bởi danh từ /cụm danh từ.

Đây là một mẫu đơn giản và rất hữu ích để trích đặc trưng. Nó là mẫu xác định cho đánh giá sản phẩm và các bài viết của diễn đàn. Mọi người thường thể hiện những bình luận hay những quan điểm trên các đặc trưng bởi mẫu đơn giản này. Ví dụ, trong dữ liệu về đệm, mọi người thường nói rằng *"no noise"* và *"no indentation"*.Ở đây *"noise"* và *"indentation"* là các đặc trưng của *mattress*. Các tác giả phát hiện rằng mẫu này thường xuyên được sử dụng trong ngữ liệu và là chỉ dẫn rất tốt cho các đặc trưng với độ chính xác cao.Nhưng các tác giả phải quan tâm đến một số thể hiện "No" cố định, như *"no problem", "no offtense"*. Trong các trường hợp này, *"problem"* và *"offense"* không phải là các đặc trưng

mong đợi. Các tác giả có một danh sách được làm bằng tay các từ này.

2.3.4 Đồ thị hai nhánh và thuật toán HITS

Tìm kiếm chủ đề bao gồm siêu liên kết (HITS) là một thuật toán phân tích liên kết để đánh giá những trang web. Các tác giả có thể áp dụng các thuật toán HITS để tính toán liên quan đặc trưng cho sắp xếp.

• Kịch bản áp dụng HITS như sau:

Đầu tiên cho một giới thiệu ngắn gọn tới HITS. Cho một truy vấn tìm kiếm rộng *q*, HITS gửi các truy vấn tới hệ thống tìm kiếm, và sau đó tập hợp *k* (*k* = 200 trong tài liệu gốc) các trang được sắp xếp cao nhất, chúng được giả định là có liên quan đến truy vấn tìm kiếm. Tập hợp này được gọi là tập gốc *R*, sau đó nó phát triển *R* bằng cách bao gồm bất kỳ trang nào trỏ vào một trang trong *R*, sau đó hình thành một tập cơ sở *S*. HITS làm việc trên các trang web trong *S*. Nó gán cho mỗi trang *S* một số điểm quyền hạn (**authority score**)và điểm trung tâm (**hub score**). Cho biết số lượng trang phải được nghiên cứu là *n*. Họ sử dụng *G = (V, E)* để biểu thị đồ thị liên kết (định hướng) của *S*. *V* là tập hợp các trang (hoặc các nút) và *E* là tập hợp định hướng các cạnh (hoặc liên kết). Họ sử dụng *L* để biểu thị ma trận kề của đồ thị.

$$L_{ij} = \begin{cases} 1 & if\ (i,j) \in E \\ 0 & otherwise \end{cases} \qquad (1)$$

Cho điểm quyền hạn của trang *i* là *A(i)*, và điểm trung tâm của trang *i* là *H(i)*. Mối quan hệ tăng cường lẫn nhau của hai điểm đại diện như sau:

$$A(i) = \sum_{(j,i)\in E} H(j) \qquad (2)$$

$$H(i) = \sum_{(i,j)\in E} A(j) \qquad (3)$$

Chúng ta có thể viết chúng dưới dạng ma trận. Họ sử dụng **A** để biểu thị vector cột với tất cả các điểm quyền hạn, **A** = (A(1), A(2),. .., A (n))^T, và sử dụng **H** để biểu thị vector cột với tất cả các điểm trung tâm, **H** = (H(1), H(2),. .., H(n))^T,

$$\mathbf{A} = L^T \mathbf{H} \quad (4)$$
$$\mathbf{H} = L\mathbf{A} \quad (5)$$

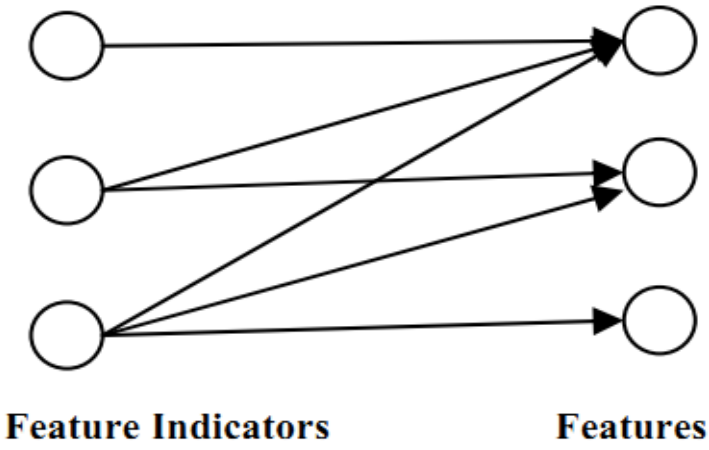
Để giải quyết vấn đề, sử dụng phương pháp lặp, bắt đầu với một số giá trị ngẫu nhiên cho các vectơ, ví dụ như, $A_0 = H_0 = (1, 1, 1,. .. 1)$. Sau đó nó tiếp tục tính toán lặp đi lặp lại cho đến khi hội tụ thuật toán.

Từ các công thức, chúng ta có thể thấy rằng điểm quyền hạn ước lượng tầm quan trọng nội dung của trang, và số điểm trung tâm ước lượng giá trị của các liên kết của nó đến các trang khác. Một điểm quyền hạn được tính toán như tổng của các điểm trung tâm tỉ lệ mà nó trỏ đến trang đó. Một điểm trung tâm được tính như là tổng của các điểm quyền hạn tỉ lệ của các trang nó trỏ tới. Ý tưởng chính của HITS là một điểm trung tâm tốt trỏ vào nhiều điểm quyền hạn tốt và điểm quyền hạn được chỉ bởi nhiều điểm trung tâm tốt. Vì vậy, quyền hạn và trung tâm có mối quan hệ tăng cường lẫn nhau.

Cho kịch bản này, các tác giả có 3 đầu mối mạnh mẽ cho các đặc trưng trong một ngữ liệu: các từ quan điểm, các mẫu bộ phận – toàn bộ, và mẫu "No". Mặc dù ba mẫu này không phải là các luật cố định, các mối quan hệ bắt buộc lẫn nhau tồn tại giữa chúng. Nếu một tính từ bổ nghĩa cho một số đặc trưng, nó có khả năng cao là một từ quan điểm tốt. Nếu một ứng cử viên đặc trưng được bổ nghĩa bởi một số từ quan điểm, nó có khả năng là một đặc trưng xác thực. Tương tự với các mẫu bộ phận – toàn bộ, mẫu "No", hoặc sự kết hợp cho ba đầu mối này. Dạng này của mối quan hệ bắt buộc lẫn nhau có thể mô hình tự nhiên trong quá trình HITS.

Áp dụng các thuật toán HITS: Dựa trên ý tưởng chính của thuật toán HITS và các chỉ dẫn đặc trưng, các tác giả áp dụng các thuật toán HITS để có được sự sắp xếp thích hợp đặc trưng. Các đặc trưng đóng vai những điểm quyền hạn và các chỉ dẫn đặc trưng đóng vai trò như những điểm trung tâm. Khác với thuật toán HITS chung, các đặc trưng chỉ có điểm quyền hạn và các chỉ dẫn đặc trưng chỉ có điểm trung tâm. Họ hình thành một đồ thị hai nhánh được định hướng, được minh họa trong Hình 2. Có thể chạy thuật toán HITS trên đồ thị hai

nhánh. Ý tưởng cơ bản nếu một ứng cử viên đặc trưng có điểm quyền hạn cao, nó phải là một đặc trưng có liên quan cao. Nếu một chỉ dẫn đặc trưng có một điểm trung tâm cao, nó phải được một chỉ dẫn đặc trưng tốt.



Hình 2: Mối quan hệ giữa các chỉ số đặc trưng và các đặc trưng

2.3.5 Sắp xếp đặc trưng

Mặc dù các thuật toán HITS có thể sắp xếp các đặc trưng dựa vào sự thích hợp đặc trưng, nhưng sắp xếp cuối cùng không chỉ được xác định dựa vào sự thích hợp. Tần suất đặc trưng là một nhân tố quan trọng khác ảnh hưởng đến thứ hạng cuối cùng. Mong muốn cao để sắp xếp chúng chính xác và các đặc trưng thường xuyên nằm ở trên, vì chúng quan trọng hơn so với các đặc trưng hiếm khi xảy ra trong khai thác quan điểm (hoặc thậm chí cả các ứng dụng khác). Với ý kiến này, các tác giả đặt tất cả chúng cùng với nhau để trình bày thuật toán mà họ sử dụng. Họ sử dụng hai bước:

Bước 1: Tính toán điểm đặc trưng sử dụng HITS mà không cần xem xét tần suất. Khởi tạo, họ sử dụng ba chỉ dẫn đặc trưng để xác định các ứng cử viên đặc trưng, từ đó tạo thành một đồ thị hai nhánh. Mỗi ứng cử viên đặc trưng đóng vai trò như là một nút quyền hạn trong đồ thị, mỗi chỉ dẫn đặc trưng đóng vai trò như một nút trung tâm. Đối với nút *s* trong đồ thị, họ cho H_S là điểm trung tâm và A_S là điểm quyền hạn. Sau đó, họ khởi tạo H_S và A_S từ 1 đến tất cả các nút trong đồ thị. Họ cập nhật các điểm của H_S và $A_S$cho đến khi chúng hội tụ bằng

cách sử dụng vòng lặp. Cuối cùng, họ chuẩn hóa A_S và tính toán điểm *S* cho một đặc trưng.

Bước 2: Hàm điểm cuối cùng xem xét tần suất đặc trưng được đưa ra trong phương trình (6).

$$S = S(f)\log(freq(f)) \qquad (6)$$

Trong đó *freq(f)* là đếm tần suất của đặc trưng *f*, và *S(f)* là số điểm quyền hạn của ứng cử viên đặc trưng *f*. Ý tưởng là đẩy vào các đặc trưng ứng cử viên thường xuyên bằng cách nhân với logarit của tần suất. Giá trị logarit được sử dụng để giảm ảnh hưởng của các số đếm tần suất lớn.

2.4 Kết quả và thảo luận

2.4.1 Tập dữ liệu

Họ sử dụng bốn tập dữ liệu khác nhau để đánh giá các kỹ thuật đề xuất. Chúng được thu từ một công ty thương mại cung cấp các dịch vụ khai thác quan điểm. Bảng 1 cho thấy các lĩnh vực và số câu trong mỗi bộ dữ liệu. Dữ liệu trong "Cars" và "Mattress" là các nhận xét sản phẩm được trích từ một số các trang web nhận xét trực tuyến. "Phone" và "LCD" là bài thảo luận diễn đàn rút từ một số trang web diễn đàn trực tuyến. Họ chia nhỏ mỗi nhận xét/bài viết thành câu và câu được được gán nhãn từ loại bằng cách sử dụng Brill's tagger (Brill, 1995). Các câu được dán nhãn là đầu vào cho hệ thống của các tác giả.

Các tập dữ liệu	**Cars**	**Mattress**	**Phone**	**LCD**
Số câu	2223	13233	15168	1783

Bảng 2. Thử nghiệm tập dữ liệu

2.4.2 Đánh giá số liệu

Bên cạnh độ chính xác và hồi tưởng, họ áp dụng độ chính xác số liệu (precision@N metric) cho đánh giá thử nghiệm (Liu, 2006). Nó cung cấp tỷ lệ

các đặc trưng chính xác nằm trong số N các ứng cử viên đặc trưng hàng đầu trong một danh sách sắp xếp. Họ so sánh các phương pháp của họ với kết quả của phương pháp double propagation mà các ứng cử viên được trích chỉ bởi tần xuất xuất hiện.

2.4.3 Kết quả thử nghiệm

Đầu tiên họ so sánh kết quả của họ với double propagation trên sự hồi tưởng và độ chính xác cho những kích thước kho ngữ liệu khác nhau. Kết quả được trình bày trong Bảng 3, 4, và 5 cho bốn tập hợp dữ liệu. Họ đưa ra độ chính xác và hồi tưởng 1000, 2000, và 3000 câu từ các tập dữ liệu. Họ đã không thử nhiều hơn bởi vì tự kiểm tra bằng tay sự hồi tưởng và chính xác trở nên khó khăn hơn. Có ít hơn 3000 câu cho các tập dữ liệu "Cars" và "LCD". Vì vậy, các cột cho "Cars" và "LCD" rỗng trong Bảng 5.

Trong bảng, "DP" đại diện cho phương pháp double propagation; "Ours" đại diện cho phương pháp đề xuất của họ; "Pr" đại diện cho chính xác, và "Re" đại diện sự hồi tưởng.

	Cars		**Mattress**		**Phone**		**LCD**	
	Pr	**Re**	**Pr**	**Re**	**Pr**	**Re**	**Pr**	**Re**
DP	0.79	0.55	0.79	0.54	0.69	0.23	0.68	0.43
Ours	0.78	0.56	0.77	0.64	0.68	0.44	0.66	0.55

Bảng 3. Kết quả của 1000 câu

	Cars		**Mattress**		**Phone**		**LCD**	
	Pr	**Re**	**Pr**	**Re**	**Pr**	**Re**	**Pr**	**Re**
DP	0.70	0.65	0.70	0.58	0.67	0.42	0.64	0.52
Ours	0.66	0.69	0.70	0.66	0.70	0.50	0.62	0.56

Bảng 4. Kết quả của 2000 câu

	Cars	Mattress		Phone		LCD
		Pr	**Re**	**Pr**	**Re**	
DP		0.65	0.59	0.64	0.48	
Ours		0.66	0.67	0.62	0.51	

Bảng 5. Kết quả của 3000 câu

Từ các bảng, chúng ta có thể thấy rằng trong tất cả các miền, phương pháp của họ thực hiện tốt hơn double propagation về hồi tưởng với một mất mát nhỏ trong sự chính xác. Trong các tập dữ liệu cho "Phone" và "Mattress", độ chính xác thậm chí còn tốt hơn. Họ cũng thấy rằng với sự gia tăng kích thước dữ liệu, khoảng cách độ hồi tưởng giữa hai phương pháp trở nên nhỏ dần và độ chính xác của cả hai phương pháp cũng giảm. Tuy nhiên, trong trường hợp này, sắp xếp đặc trưng đóng một vai trò quan trọng trong phát hiện các đặc trưng quan trọng.

So sánh sắp xếp giữa hai phương pháp được thể hiện trong Bảng 6, 7, và 8, trong đó cung cấp độ chính xác của kết quả trên 50, 100 và 200 tương ứng. Lưu ý rằng các thực nghiệm báo cáo trong các bảng này đã được chạy trên toàn bộ các tập dữ liệu. Không có nhiều kết quả hơn cho dữ liệu "LCD" vượt ra ngoài tốp 200 như là chỉ có một số giới hạn các đặc trưng được thảo luận trong dữ liệu. Vì vậy, các cột "LCD" trong Bảng 7 là rỗng. Họ sắp xếp các ứng cử viên đặc trưng được trích dựa trên tần suất cho phương pháp double propagation (DP). Sử dụng tần suất xuất hiện như là cách tự nhiên để sắp xếp các đặc trưng. Một đặc trưng thường xuyên xuất hiện trong một kho ngữ liệu, thì nó là quan trọng hơn. Tuy nhiên, sắp xếp dựa trên tần suất giả thuyết các ứng cử viên được trích là các đặc trưng chính xác. Các bảng cho thấy rằng phương pháp đề xuất của họ (Ours) nhanh hơn so với double propagation đáng kể. Lý do là một số ứng cử viên đặc trưng có tần suất xuất hiện thường xuyên cao được trích bởi double propagation không phải là các đặc trưng đúng. Phương pháp của họ xem xét liên quan đặc trưng là một nhân tố quan trọng. Vì vậy, nó cho ra các sắp xếp

tốt hơn.

	Cars	Mattress	Phone	LCD
DP	0.84	0.81	0.64	0.68
Ours	0.94	0.90	0.76	0.76

Bảng 6. Độ chính xác ở top 50

	Cars	Mattress	Phone	LCD
DP	0.82	0.80	0.65	0.68
Ours	0.88	0.85	0.75	0.73

Bảng 7. Độ chính xác ở top 100

	Cars	Mattress	Phone	LCD
DP	0.75	0.71	0.70	
Ours	0.80	0.79	0.76	

Bảng 8. Độ chính xác ở top 200

CHƯƠNG 3: THỰC NGHIỆM

3.1 Công cụ gán nhãn từ loại Stanford Parser

3.1.1 Giới thiệu

Phân tích cú pháp ngôn ngữ tự nhiên là một chương trình hoạt động dựa vào cấu trúc ngữ pháp của câu, ví dụ, với nhóm từ đi cùng nhau (như "phrases") và với từ là chủ đề hoặc đối tượng của động từ. Xác xuất phân tích cú pháp sử dụng kiến thức về ngôn ngữ có được từ phân tích câu thủ công để tìm cách sản suất phân tích có khả năng nhất của những câu mới. Thống kê những phân tích cú pháp này vẫn còn một số sai lầm, nhưng thường làm việc khá tốt. Phát triển của họ là một trong những bước đột phá lớn nhất trong việc xử lý ngôn ngữ tự nhiên trong những năm 1990.

Stanford biểu diễn kiểu phụ thuộc được thiết kế để cung cấp một mô tả đơn giản của các mối quan hệ ngữ pháp trong một câu có thể dễ dàng hiểu và hiệu quả được sử dụng bởi những người không có chuyên môn ngôn ngữ học, những người muốn trích xuất văn bản quan hệ.

Gói sản phẩm này là một cài đặt Java của xác suất phân tích cú pháp ngôn ngữ tự nhiên, cả PCFG tối ưu hoá cao và bộ phân tích cú pháp từ vựng phụ thuộc, và phân tích cú pháp từ vựng PCFG. Các phiên bản gốc của phân tích cú pháp này chủ yếu được viết bởi Dan Klein, với mã hỗ trợ và phát triển ngữ pháp ngôn ngữ Christopher Manning. Mở rộng thêm công việc (quốc tế và mẫu ngôn ngữ cụ thể, đầu vào / đầu ra linh hoạt, ngữ pháp nén chặt, mạng tinh thể phân tích cú pháp, k-phân tích cú pháp tốt nhất, đánh máy phụ thuộc đầu ra, hỗ trợ người dùng, vv) đã được thực hiện bởi Roger Levy, Christopher Manning, TeG Grenager, Galen Andrew, Marie-Catherine de Marneffe, Bill MacCartney, Anna Rafferty, Spence Green, Huihsin Tseng, Pi-Chuan Chang, Wolfgang Maier, và Jenny Finkel.

Phiên bản hiện tại của phân tích cú pháp yêu cầu Java 6 (JDK1.6) hoặc những phiên bản sau. (Bạn cũng có thể tải về một phiên bản cũ của phân tích cú

pháp, phiên bản 1.4, chạy theo JDK 1.4 hoặc phiên bản 2.0 mà chạy theo JDK 1.5, tuy nhiên hiện nay những phiên cũ ít được hỗ trợ). Phân tích cú pháp cũng đòi hỏi dung lượng hợp lý của bộ nhớ (tại ít nhất là 100MB để chạy như là một phân tích cú pháp PCFG trên câu lên đến 40 từ trong chiều dài, thường khoảng 500MB bộ nhớ để có thể phân tích cú pháp tương tự điển hình-của-Newswire câu bằng cách sử dụng mẫu yếu tố).

3.1.2 Cách sử dụng

Công cụ gán nhãn từ loại Stanford là thiết kế được sử dụng từ dòng lệnh hoặc lập trình thông qua giao diện ứng dụng của nó.

Có thể sử dụng phương pháp sau:

1. Trên hệ thống Windows, bạn có thể chạy một giao diện phân tích cú pháp bằng cách nhấp đúp vào biểu tượng lexparser-gui.bat, hoặc đưa ra các lệnh lexparser-gui trong thư mục này từ một dấu nhắc lệnh.

   ✓ Nhấp Load File, Browse, và điều hướng đến và chọn testsent.txt

   ✓ Nhấp Load Parser, Browse và chọn thư mục tương tự một mẫu jar. Từ các mẫu jar chọn englishPCFG.ser.gz.

   ✓ Nhấp Parser để bắt đầu phân tích câu.

2. Trên hệ thống Ubuntu, đưa ra các lệnh lexparser trong thư mục này từ một dấu nhắc lệnh: sh lexparser.sh file_dữ_liệu_nguồn > file_đích

3.2 Chương trình thực nghiệm

3.2.1 Bài toán

Input: Cho một tập hợp các câu văn bản đánh giá có quan điểm về sản phẩm hoặc đối tượng.

Output: Tập từ quan điểm và các đặc trưng được trích chọn và sắp xếp.

3.2.1.1 Bộ dữ liệu

Chúng tôi sử dụng bộ dữ liệu 500 nhận xét để trích các đặc trưng và quan điểm về các đặc trưng sản phẩm.

Trước khi tiến hành thực nghiệm, chúng tôi sử dụng công cụ gán nhãn từ loại Stanford Parser để lấy thông tin từ loại cho các câu.

3.2.1.2 Phương pháp

Khai thác đặc trưng cho các thực thể là một nhiệm vụ quan trọng trong nhiệm vụ khai thác quan điểm.

Thuật toán gồm 4 bước:

- Gán nhãn từ loại bằng công cụ stanford parser
- Xác định câu có chứa một hay nhiều dấu hiệu sản phẩm hay từ quan điểm được xem là câu chỉ quan điểm
- Trích chọn đặc trưng

✓ Với mỗi câu trong dữ liệu chỉ quan điểm, rút tất cả những danh từ / cụm danh từ được coi là những từ chỉ đặc trưng và những tính từ được coi là những từ chỉ quan điểm.Các quan điểm và các đặc trưng đã được trích được sử dụng để xác định những quan điểm và đặc trưng mới. Quá trình này cứ lặp đi lặp lại cho đến khi không thể tìm thấy các từ quan điểm hay đặc trưng thì kết thúc.

✓ Dựa vào mối quan hệ ngữ nghĩa giữa quan điểm và đặc trưng để xác định đặc trưng và quan điểm trong dữ liệu. Áp dụng bộ phân tích cú pháp phụ thuộc dựa vào ngữ pháp phụ thuộc. Sử dụng các luật trong Qiu và các cộng sự, 2009:

- $R1_i$ để trích các từ quan điểm (s) sử dụng các từ quan điểm words (S_i)

- $R2_i$ để trích các từ quan điểm (s) sử dụng các đặc trưng (F)

- $R3_i$ để trích các đặc trưng (f) sử dụng các từ quan điểm (S)

- $R4_i$ để trích các đặc trưng (f) sử dụng các đặc trưng đã được trích (F_i).

- Sắp xếp đặc trưng sản phẩm áp dụng giải thuật HITS

✓ Các đặc trưng đóng vai những điểm quyền hạn và các chỉ số đặc trưng đóng vai những điểm trung tâm. Khác nhau từ thuật toán HITS chung, các đặc trưng có điểm quyền hạn và đặc trưng có điểm trung tâm trong trường hợp của họ.

✓ Hình thành một đồ thị hai nhánh được định hướng. Chạy các thuật toán HITS trên đồ thị hai nhánh. Ý tưởng cơ bản nếu một ứng cử viên đặc trưng có điểm quyền hạn cao, nó phải là một đặc trưng có liên quan. Nếu một đặc trưng chỉ có một số điểm trung tâm cao, nó phải được một chỉ số đặc trưng tốt.

3.2.2 Thực nghiệm

Các luật, các mẫu dựa trên mối quan hệ giữa hai từ được sử dụng để trích từ quan điểm hay đặc trưng:

- conjunctions (word1_JJ, word2_JJ).

Vd: *Here 's the brief synopsis : the phone is tiny , cute , feels kind of " plastic-like " (as if it might break) , but seems pretty sturdy*.

conj_but(tiny_JJ, sturdy_JJ).

- {MR} (word1_JJ/NN, word2_JJ/NN). Trong đó: {MR} bao gồm các mối quan hệ phụ thuộc mô tả các mối quan hệ, thí dụ như: mod, subj, obj, ...

✓ Vd: *I am a business user who heavily depend on mobile service .*

amod(service_NN, mobile_JJ).

✓ Vd: *After years with that carrier 's expensive plans and horrible customer service , portability seemed heaven-sent .*

amod(plans_NNS, expensive_JJ);

amod(service_NN, horrible_JJ);

nsubj(heaven-sent_NN, portability_NN).

- conjunctions (word1_NN, word2_NN).

Vd: *My favorite features , although there are many , are the speaker phone , the radio and the infrared .*

conj_and(phone_NN, radio_NN);

conj_and(phone_NN, infrared_NN).

- NP + Prep + CP : danh từ / cụm từ danh từ (NP) chứa đựng từ bộ phận và cụm khái niệm lớp (CP) có chứa từ toàn bộ. Chúng được nối bởi giới từ (Prep), thí dụ, ***"of", "in", "on", "about"***,….

Vd: *There is much which has been said in other reviews about the features of this phone , it is a great phone , mine worked without any problems right out of the box .*

prep_in(said, reviews);

prep_about(said, features);

prep_of(features, phone);

prep_without(worked, problems); …

- CP + with + NP: CP là cụm khái niệm lớp, và NP là cụm danh từ / danh từ. Chúng được nối với nhau bằng từ ***"with"***.

Vd: *The speaker phone is very functional and i use it in the car , very audible even with freeway noise*.

prep_with(audible_JJ, noise_NN).

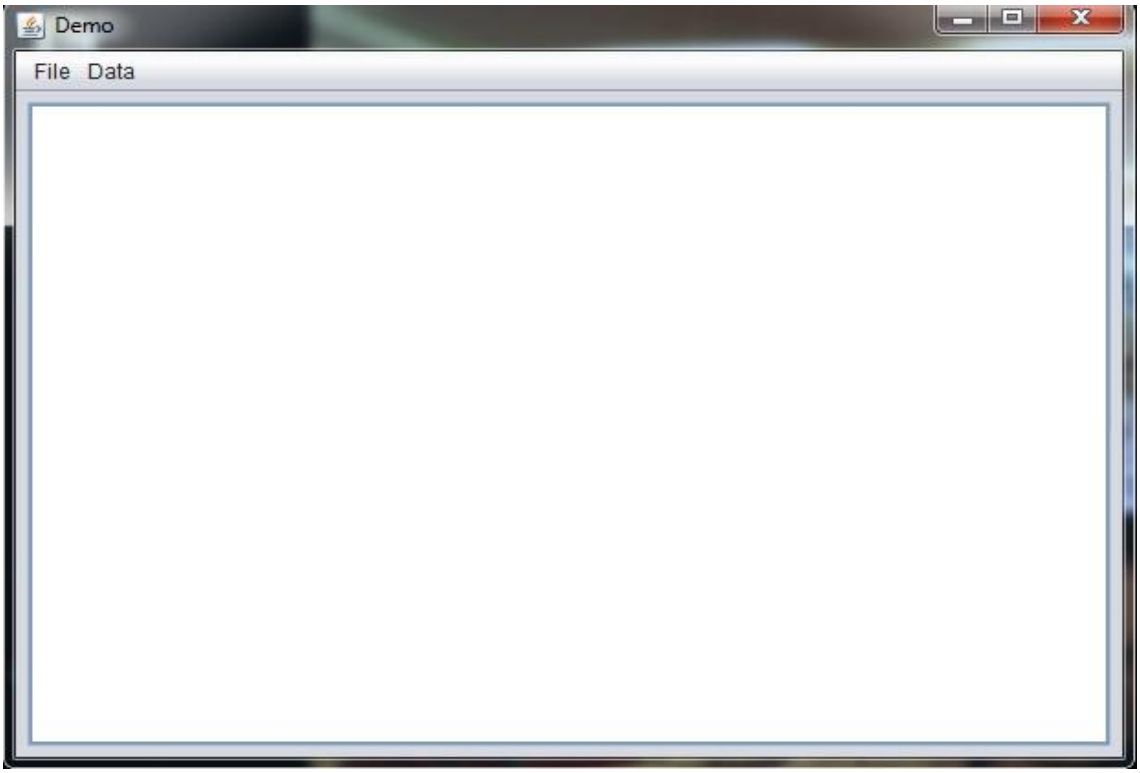
- NP CP hay CP NP: danh từ / cụm danh từ (NP) và cụm khái niệm lớp (CP) tạo thành một từ ghép. Ví dụ, *"mattress pad"*. *"Pad"* là một đặc trưng của *"mattress"*.
- CP Verb NP: CP là cụm khái niệm lớp có chứa từ toàn bộ, NP là cụm danh từ có chứa các từ bộ phận và động từ có giới hạn và xác

định. Sử dụng động từ thể hiện để tìm các quan hệ bộ phận - toàn bộ trong một câu, thí dụ, *"has", "have" "include" "contain" "consist", "comprise",…*

3.2.3 Kết quả

Chương trình thử nghiệm trích được các đặc trưng sản phẩm và các nhận xét cho các đặc trưng đó dựa theo các luật và mẫu bộ phận – toàn bộ.

Giao diện chính của chương trình



Mở file dữ liệu:

Trích đặc trưng:

```
F:\Doan2012\Thanh\NewData\nokia66par1.txt
 phone,excellent

 service,excellent

 user,i

 user,i

 depend,user

 depend,user

 service,mobile
```

KẾT LUẬN

Trong quá trình thực hiện đồ án, em đã tìm hiểu về phân tích quan điểm hay khai thác quan điểm và các vấn đề đặt ra với bài toán này. Tìm hiểu kỹ về phương pháp trích và sắp xếp đặc trưng sản phẩm trong tài liệu chứa quan điểm.

Đồ án cũng đã đi tìm hiểu các đặc trưng ngôn ngữ dùng cho bài toán trích đặc trưng như: phân tích cú pháp, gán nhãn từ loại cho ngôn ngữ; Với một số phương pháp trích và sắp xếp đặc trưng sản phẩm trong tài liệu chứa quan điểm.

Chương trình thực nghiệm cũng đã thử nghiệm trên bộ dữ liệu 500 nhận xét để trích các đặc trưng và quan điểm về các đặc trưng sản phẩm.

Hướng nghiên cứu tiếp theo, em sẽ tiếp tục thử nghiệm với các phương pháp trích đặc trưng khác và nghiên cứu các phương pháp sắp xếp đặc trưng để sắp xếp đặc trưng hiệu quả.

Trong một khoảng thời gian có hạn, nên khi trình bày các vấn đề em đã

nghiên cứu được không tránh khỏi những thiếu sót. Em rất mong nhận được những ý kiến đóng góp quý báu của thầy cô và các bạn.

Em xin chân thành cảm ơn!

TÀI LIỆU THAM KHẢO

Tiếng Việt:

1. Ths. Nguyễn Thị Xuân Hương và Ths. Lê Thuy, Phân tích quan điểm và một số tiếp cận, Hội nghị khoa học CNTT lần thứ nhất, 2012.

2. Nguyễn Mạnh Đức, Tìm hiểu về khai thác quan điểm và phân loại quan điểm ở mức câu, Khoá luận tốt nghiệp hệ đại học ngành Công nghệ thông tin, Đại học Dân lập Hải Phòng, 2011.

Tiếng Anh:

1. Extracting and Ranking Product Features in Opinion Documents. Lei Zhang, Bing Liu, Suk Hwan Lim and Eamonn O'Brien-Strai, 2010

2. Expanting Domain Sentiment Lexicon through Double Propagation. Guang Qiu, Bing Liu, Jiajun Bu and Chun Chen, 2009

3. http://nlp.stanford.edu/software/lex-parser.shtml

4. http://www.cs.uic.edu/~liub/FBS/FBS.html

5. http://nlp.stanford.edu/software/stanford-dependencie s.shtml