

บทคัดย่อ

กูเกิลเพจแรงค์ (Google's PageRank) เป็นอัลกอริทึมที่สำคัญอันหนึ่งในการทำเครื่องจักรค้นหาเว็บ (Web search engine) โดยในกูเกิลเพจแรงค์ เว็บเพจทั้งหมดจะถูกแทนด้วยเว็บลิงก์กราฟ (Web-link graph) โดยโหนดของกราฟจะใช้แทนเว็บเพจ และเอ็ดจ์ของกราฟจะใช้แทนเว็บลิงค์ การคำนวณเพจแรงค์จะทำงานได้อย่างไม่ค่อยมีประสิทธิภาพหากเว็บลิงก์กราฟมีขนาดใหญ่มาและมีการเพิ่มขนาดเว็บลิงก์กราฟ เนื่องจากต้องทำการคำนวณใหม่ทั้งหมด เพื่อแก้ปัญหาดังกล่าว จึงมีผู้นำเสนอแนวคิดการคำนวณเพจแรงค์บนระบบเครือข่ายแบบเพียร์ทูเพียร์ โดยในแต่ละเพียร์จะเก็บเว็บลิงก์กราฟไว้บางส่วน ซึ่งเว็บลิงก์กราฟที่ถูกเก็บไว้ในแต่ละเพียร์ก็คือกราฟย่อยของเว็บลิงก์กราฟขนาดใหญ่ แล้วแต่ละเพียร์จะแยกกันคำนวณเพจแรงค์แบบกระจายศูนย์ จำนวนรอบที่ใช้ในการคำนวณเพจแรงค์จะขึ้นอยู่กับความหนาแน่นของกราฟย่อยที่ถูกเก็บไว้ในแต่ละเพียร์ โดยหากกราฟย่อยมีความหนาแน่นสูง จะทำให้จำนวนรอบในการคำนวณเพจแรงค์ก็จะสูงตามไปด้วย และจะส่งผลถึงการคำนวณเพจแรงค์ของเครือข่ายทั้งหมด เพราะในการคำนวณเพจแรงค์บนเครือข่ายเพียร์ทูเพียร์นั้นจะนับเวลาในการคำนวณเพจแรงค์จากเพียร์ที่ทำงานช้าที่สุด การทำให้เว็บลิงก์กราฟที่ถูกเก็บไว้ในแต่ละเพียร์มีค่าความหนาแน่นใกล้เคียงกันจะช่วยลดเวลาของการคำนวณเพจแรงค์บนเครือข่ายแบบกระจายศูนย์ลงได้ และเมื่อมีการเพิ่มขนาดเว็บลิงก์กราฟ และต้องทำการคำนวณเพจแรงค์ใหม่ ในงานวิจัยนี้จึงได้เสนอปัญหาแบ่งกราฟแบบความหนาแน่นสมดุลขึ้น เพื่อให้การคำนวณเพจแรงค์บนระบบแบบกระจายศูนย์มีประสิทธิภาพมากยิ่งขึ้น โดยสามารถใช้ในการแบ่งกราฟทั้งในการทำงานเริ่มต้น (Start-up) และการทำงานแบบเพิ่มค่า (Incremental) ซึ่งจากผลการทดลองแสดงให้เห็นว่า วิธีการที่นำเสนอทำให้สามารถแบ่งกราฟให้มีความหนาแน่นสมดุลมากขึ้น เมื่อเทียบกับการแบ่งแบบดั้งเดิม

Abstract

Google's PageRank is the most notable approach for web search ranking. In general, web pages are represented by web-link graph; a web-page is represented by a node, and a link between two pages is represented by an edge. In particular, it is not efficient both to perform PageRank of a large web-link graph in a single computer and when new web pages are added to the web-link graph. Distributed systems, such as P2P, are viable choices to address such limitation. In P2P-based PageRank, each computational peer contains a partial web-link graph, i.e., a sub-graph of the global web-link graph, and its PageRank is computed locally. The convergence time of a PageRank calculation is affected by the web-link graph density, i.e., the ratio of the number of edges to the number of nodes, such that if a web-link graph has high density, it will take longer time to converge. As the execution time to compute the P2P-based web ranking is influenced by the execution time of the slowest peer to compute the local ranking, the density-balanced local web-link graph partitioning can be highly desirable. This paper addresses a density-balanced partitioning problem and proposes an efficient algorithm for the problem that will benefit both for starting up and also when web pages are incrementally increased. The experiment results show that the proposed algorithm can effectively partition graph into density-balanced sub-graph with an acceptable cost.