ในปัจจุบันข้อมูลในระบบอินเทอร์เน็ตหรือเวิลด์ ไวด์ เว็บ นั้นได้เพิ่มจำนวนขึ้นเป็นจำนวนมาก ด้วยความรวดเร็ว รวมถึงข้อมูลบุคคลซึ่งเป็นสิ่งที่ผู้ใช้ต้องการค้นหามากถึง 30 % วิธีการที่เป็นที่นิยมสำหรับใช้เพื่อค้นหาข้อมูลบุคคลจากเวิลด์ ไวด์ เว็บ คือ การค้นหาข้อมูลผ่านบริการของเสิร์ชเอนจิ้น อย่างไรก็ตาม วิธีค้นหาข้อมูลในลักษณะนี้ ทำให้ผู้ใช้ต้องเสียเวลานานในการค้นหาข้อมูล โดยเวลาที่สูญเสียไปเกิดขึ้นจาก 2 สาเหตุคือ เวลาที่ใช้ในการเลือกคำค้นที่เหมาะสมเพื่อให้ได้เอกสารที่มีข้อมูลของบุคคลที่ต้องการ และเวลาที่ต้องใช้ในการตรวจสอบผลลัพธ์จากการค้นคืนของเสิร์ชเอนจิ้นว่าเป็นข้อมูลที่ผู้ใช้ต้องการหรือไม่ อีกทั้งข้อมูลบุคคลที่ต้องการอาจจะกระจัดกระจายเว็บเพจต่างๆทำให้เสียเวลาในการรวบรวมอีกด้วย ดังนั้นงานวิจัยนี้จึงวิจัยและพัฒนาระบบสืบค้นข้อมูลบุคคลจากเวิลด์ ไวด์ เว็บ โดยระบบนี้จะรวบรวมเว็บเพจที่มีข้อมูลบุคคลปรากฏอยู่แล้วสกัดเฉพาะข้อมูลส่วนที่เกี่ยวกับบุคคลเท่านั้น เช่น ชื่อ นามสกุล ตำแหน่ง องค์กรที่สังกัด โดยก่อนจัดเก็บข้อมูลที่สกัดได้จะมีการตรวจสอบชื่อของบุคคลว่าชื่อใดเป็นชื่อของบุคคลคนเดียวกันบ้าง เพื่อที่จะรวมข้อมูลเข้าด้วยกัน และเปิดบริการให้ผู้ใช้สามารถสืบค้นข้อมูลเหล่านี้ได้ทั้งทางตรงและทางอ้อม โดยสำหรับการเข้าถึงข้อมูลโดยทางอ้อมนั้นจะใช้เทคนิคอนุมานแบบอาศัยกฎ ซึ่งจะกระทำบนข้อมูลที่เก็บอยู่ในโครงสร้างแบบอาร์ดีเอฟ (RDF) เพื่อช่วยให้ผู้ใช้สามารถเข้าถึงข้อมูลเหล่านี้ได้

สิ่งที่พัฒนาขึ้นในระบบสืบค้นข้อมูลบุคคลในงานวิจัยนี้มี 3 ส่วน ได้แก่ ส่วนรวบรวมเว็บเพจที่มีข้อมูลบุคคล ส่วนรวมข้อมูลที่สกัดมาได้เข้าไว้ด้วยกัน โดยจะตรวจสอบหาชื่อคนที่คล้ายกันว่าเป็นชื่อของคนๆเดียวกันหรือไม่ ซึ่งได้พัฒนาจากวิธีการตรวจสอบความคล้ายของสตริง (Edit distance) เพิ่มเติมให้เหมาะสมกับภาษาไทย และส่วนของการให้บริการสืบค้นข้อมูล ซึ่งในส่วนนี้จะมีการเพิ่มเทคนิคอนุมานแบบอาศัยกฎ และออนโทโลจี (Ontology) ซึ่งใช้ในการจัดเก็บความรู้เกี่ยวกับตำแหน่งของบุคคล และสาขาความเชี่ยวชาญในหน่วยงานมหาวิทยาลัย เพื่อช่วยให้ผู้ใช้สามารถเข้าถึงข้อมูลที่เกิดจากการสรุปได้

เมื่อทดลองกับเว็บเพจที่รวมมาได้ทั้งสิ้นจำนวน 169,079 เว็บเพจ ผลปรากฏว่า ส่วนรวมข้อมูลบุคคลที่สกัดได้เข้าไว้ด้วยกันให้ค่าความเที่ยง 0.85 และค่าระลึกได้ 0.88 เมื่อใช้ค่าขอบเขตเท่ากับ 1.5 สำหรับส่วนของการสืบค้นข้อมูลให้ค่าความถูกต้อง 0.83 และค่าระลึกได้ 0.85 เมื่ออาศัยออนโทโลจีและกฎการอนุมานแล้วค่าความถูกต้องเพิ่มขึ้น 0.04 ค่าระลึกได้เพิ่มขึ้น 0.22

Nowadays, the amount of useful information provided by the Internet is currently growing at an incredible rate. Recently, searching for person information from the Internet is 30% of all search queries. The method usually used for searching person information from WWW is to use traditional search engine. However, users must spend a lot of time finding appropriate query words, reading retrieved documents to find required information, and, in some situations, gathering information from various sources. Therefore, this research proposed a person information searching system from WWW that automatically collects web pages having person information, and extracts person information (e.g. name, surname, position, and organization) from collected web pages. The system also provides a service for searching person information both direct and indirect way. To search for the information indirectly, the system use a rule-based reasoning technique that processes information stored in RDF model.

Three important components of person information searching system consisting of person information gathering, person information integration, and information accessing are developed in this research. A new edit distance based similarity checking method is proposed to check the similarity of Thai person name in information integration module. The information accessing component uses a reasoning technique to access required information indirectly.

The experiment with 169,079 collected web pages from university and organization shows that the precision and recall of the information integration module are 0.85 and 0.88, and the precision and recall of the information accessing module are 0.83 and 0.85 respectively. Moreover, when using ontology and reasoning technique the precision and recall are increased by 0.04 and 0.22.