مقاله‌ای که خلاصه‌ی آن در ادامه می‌آید مروری است بر ادبیات موضوع و کارهای انجام گرفته در زمینه استخراج خودکار کلمات کلیدی (automatic keyword extraction) و خلاصه‌سازی متن (text summarization). یکی از کارهایی که در خلاصه‌سازی متن اهمیت زیادی دارد و مرحله‌ی اصلی آن محسوب می‌شود، استخراج کلمات کلیدی است.

متن‌کاوی به فرایندی گفته می‌شود که در آن با استخراج مقدار زیادی متن، تلاش می‌شود تا به اطلاعات با کیفیتی رسید. در متن‌کاوی، برای تحلیل و آنالیز متن، از برخی از تکنیک‌های رایج در پردازش زبان‌های طبیعی استفاده می‌شود مانند برچسب‌زنی نقش کلمات یا اصطلاحا برچسب‌زنی جزء کلام (part-of-speech tagging)، تجزیه کردن متن (parsing)، روش n-gramها[1]، مشخص‌سازی توکن‌ها و غیره. توجه شود که متن‌کاوی شامل کارهایی مانند استخراج کلمات کلیدی و خلاصه‌سازی متن می‌شود.

استخراج خودکار کلمات کلیدی به فرایندی گفته می‌شود که در آن بدون دخالت انسان، کلمات و عباراتی از متن سند انتخاب می‌شوند که بتوانند به بهترین صورت ایده‌ی اصلی حاکم بر سند را بیان کنند.

خلاصه‌سازی متن به فرایندی گفته می‌شود که در آن مهمترین ویژگی‌ها یا اطلاعات متن، استخراج شده و به صورت چکیده‌ای از سند اصلی ارائه می‌شوند. اندازه‌ی خلاصه‌ها معمولا در حدود ۱۷٪ اندازه متن اصلی می‌باشد؛ با وجود این، خلاصه، تمام چیزی را که از خواندن متن اصلی می‌توان فهمید در خود دارد. دو روش برای خلاصه‌سازی متن وجود دارد: خلاصه‌سازی چکیده‌ای و خلاصه‌سازی استخراجی. خلاصه‌سازی چکیده‌ای مانند وقتی است که یک فرد مقاله‌ای را می‌خواند و سپس خلاصه‌ای از آن را بازگو می‌کند. هنوز الگوریتم استانداردی برای این نوع خلاصه‌سازی تهیه نشده است. خلاصه‌سازی استخراجی، خلاصه را از متن خود نوشته استخراج می‌کند.

روش‌های استخراج کلیدواژه را می‌توان به چهار دسته تقسیم نمود: روش‌های آماری ساده، روش‌هایی که بر اساس زبان‌شناسی کار می‌کنند، روش‌های یادگیری ماشینی و روش‌هایی که ترکیبی از روش‌های قبلی هستند.

- روش‌های آماری ساده: این روش‌ها معمولا داده‌ی train ندارند و معیارهایی که برای مشخص کردن کلمات کلیدی سند استفاده می‌کنند به زبان‌شناسی ربطی ندارد مانند مکان کلمه در سند یا tf-idf یک کلمه یا تعداد دفعاتی که دو کلمه پشت سر هم آمده‌اند (co-occurrence[2]).
- روش‌های مبتنی بر زبان‌شناسی: در این روش از خصوصیات زبانی کلمات برای تشخیص کلمات کلیدی استفاده می‌شود؛ با انجام کارهایی مانند تحلیل لغوی متن، تحلیل معنایی متن و غیره.
- روش‌های یادگیری ماشینی: که داده‌های train باید به صورت دستی آماده شوند. از جمله‌ی این روش‌ها می‌توان مدل پنهان مارکوف و روش بیز ساده را نام برد. همچنین یکی از معروف‌ترین الگوریتم‌ها در این زمینه الگوریتم KEA است.
- روش‌های ترکیبی: هم می‌تواند ترکیبی از روش‌های ذکر شده قبلی باشد و هم می‌تواند از روش‌های اکتشافی (هیوریستیک) استفاده کند.

در ادامه، مقالات منتشر شده در موضوع استخراج کلیدواژه و روش به کار رفته در هر مقاله بیان شده است. سپس مروری کلی بر موضوع خلاصه‌سازی متن و سابقه تحقیق در این زمینه صورت گرفته است.

---

[1] n-gram: هر n-gram شامل n قسمت پشت سر هم از متن می‌باشد که قسمت می‌تواند یک حرف، یک هجا، یک کلمه و یا غیره باشد.

[2] co-occurrence: این که دو کلمه چند بار پشت سر هم (کنار هم) و با ترتیب مورد نظر تکرار شده‌اند.

از مطالب مقاله موضوع خلاصه‌سازی مورد نظر ما نیست. در موضوع استخراج کلمات کلیدی، طبق گفته‌ی نویسندگان، روش‌های آماری ساده که در آن‌ها معیارهایی مانند مکان واژگان در متن یا tf-idf واژگان مورد نظر قرار می‌گیرد، دقت مناسبی ندارند. بنابراین به نظر می‌رسد روش‌های مبتنی بر زبان‌شناسی یا روش‌های یادگیری ماشینی یا ترکیبی از این دو برای استخراج کلمات کلیدی مناسب‌تر باشند.