Unscharfe Symbole

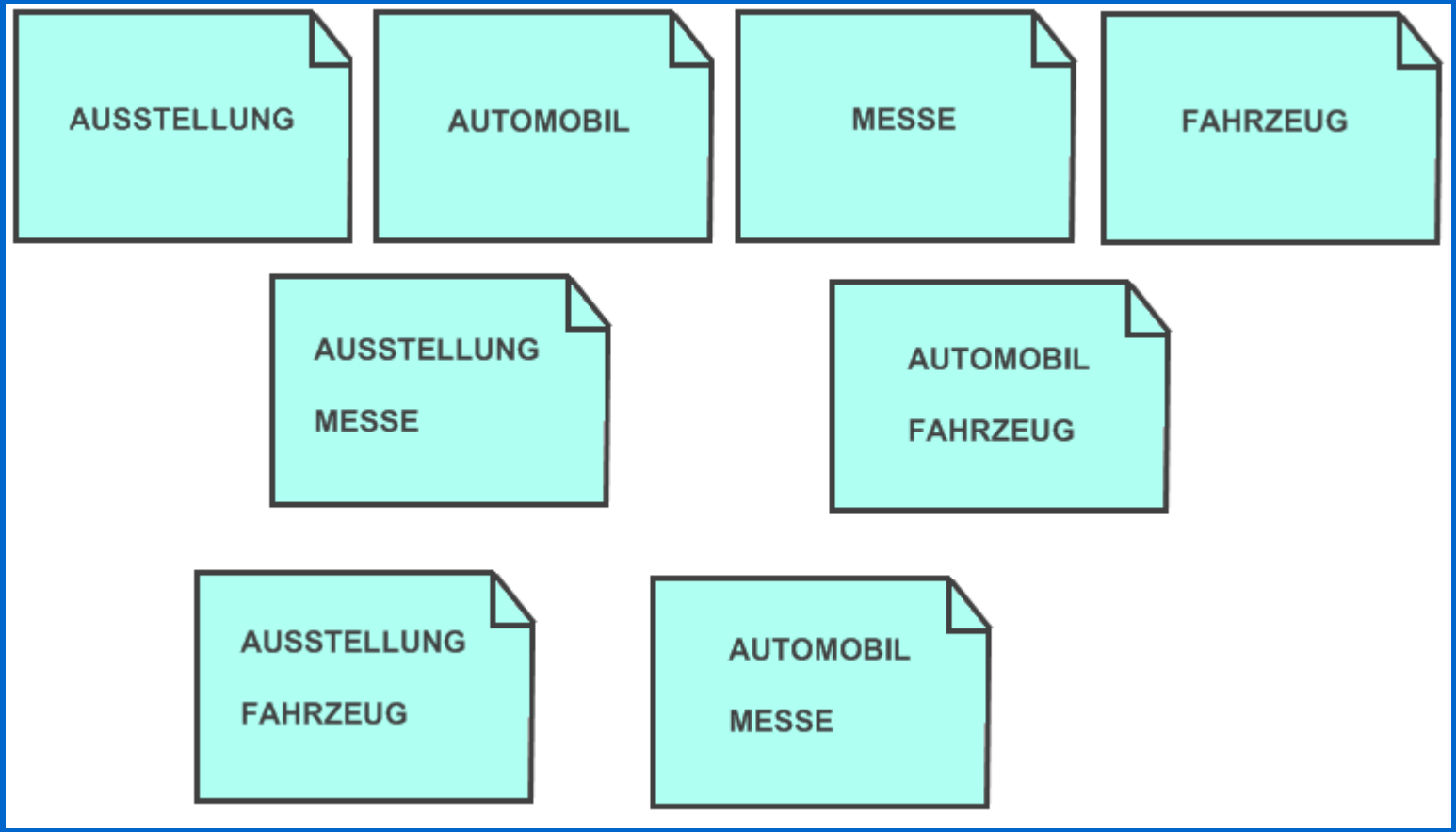
Theoretischer Hintergrund

Aufgabe: suche nach

AUSSTELLUNG AUTOMOBIL MESSE FAHRZEUG

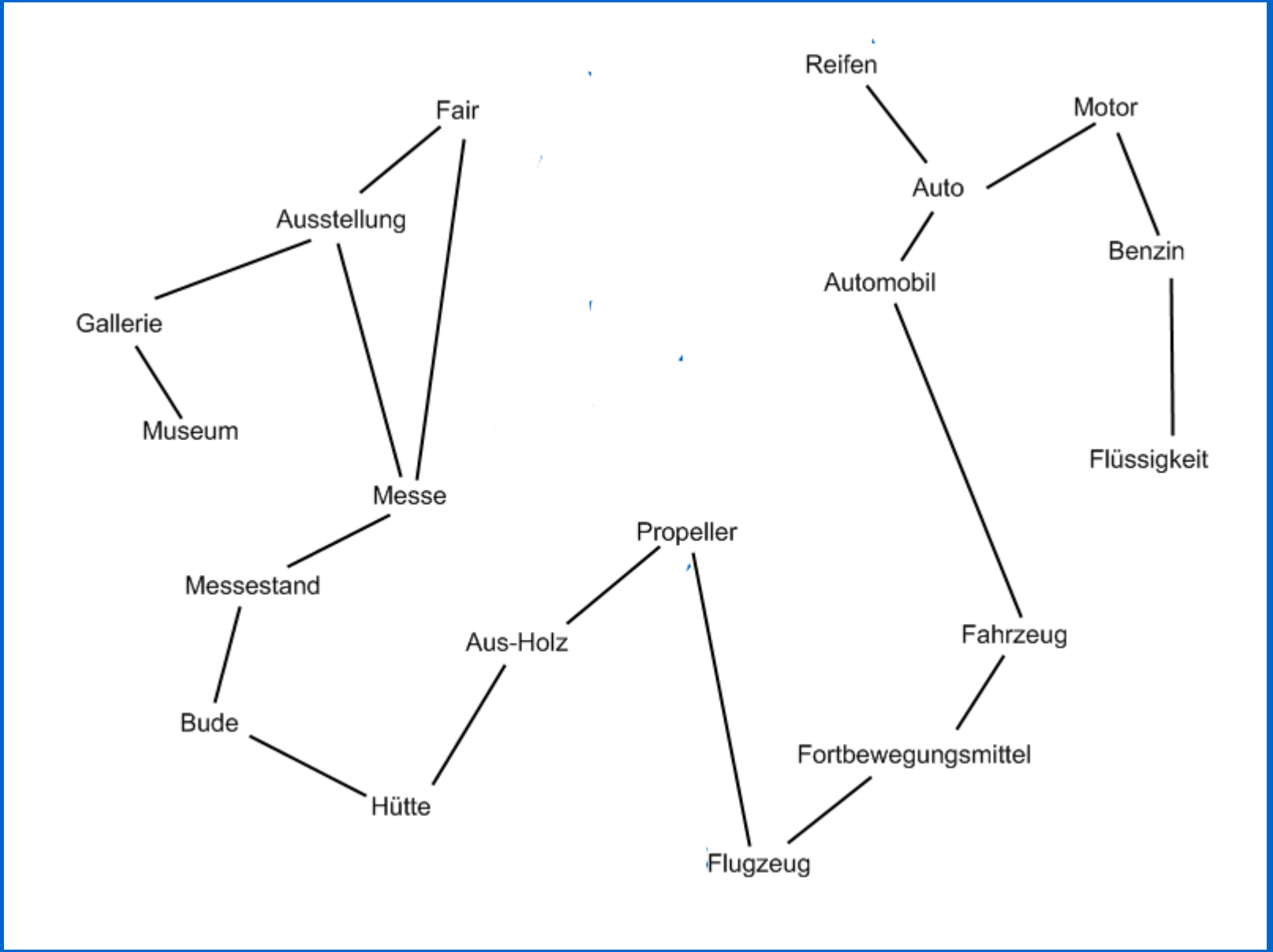
Suchbegriffe:

AUSSTELLUNG AUTOMOBIL MESSE FAHRZEUG



Intuitiv spürt man, dass die Dokumente der letzten Reihe besser zur Anfrage passen als die der vorletzten Reihe. Eine Suchmaschine sollte den semantischen Abstand der Suchbegriffe beachten.

Ein semantisches Netz liefert den Abstand zwischen je zwei Begriffen.



Schon jetzt hätte man bereits eine Verbesserung: man müsste lediglich in einem Dokument für alle Paare (S_i,S_j) von Suchbegriffen testen, ob beide Suchbegriffe im Dokument vorkommen und in diesem Falle die Werte der semantischen Distanz D(S_i,S_j) aufzuaddieren.

Je weiter zwei Begriffe voneinander entfernt sind, desto höher der Anteil am Rang.

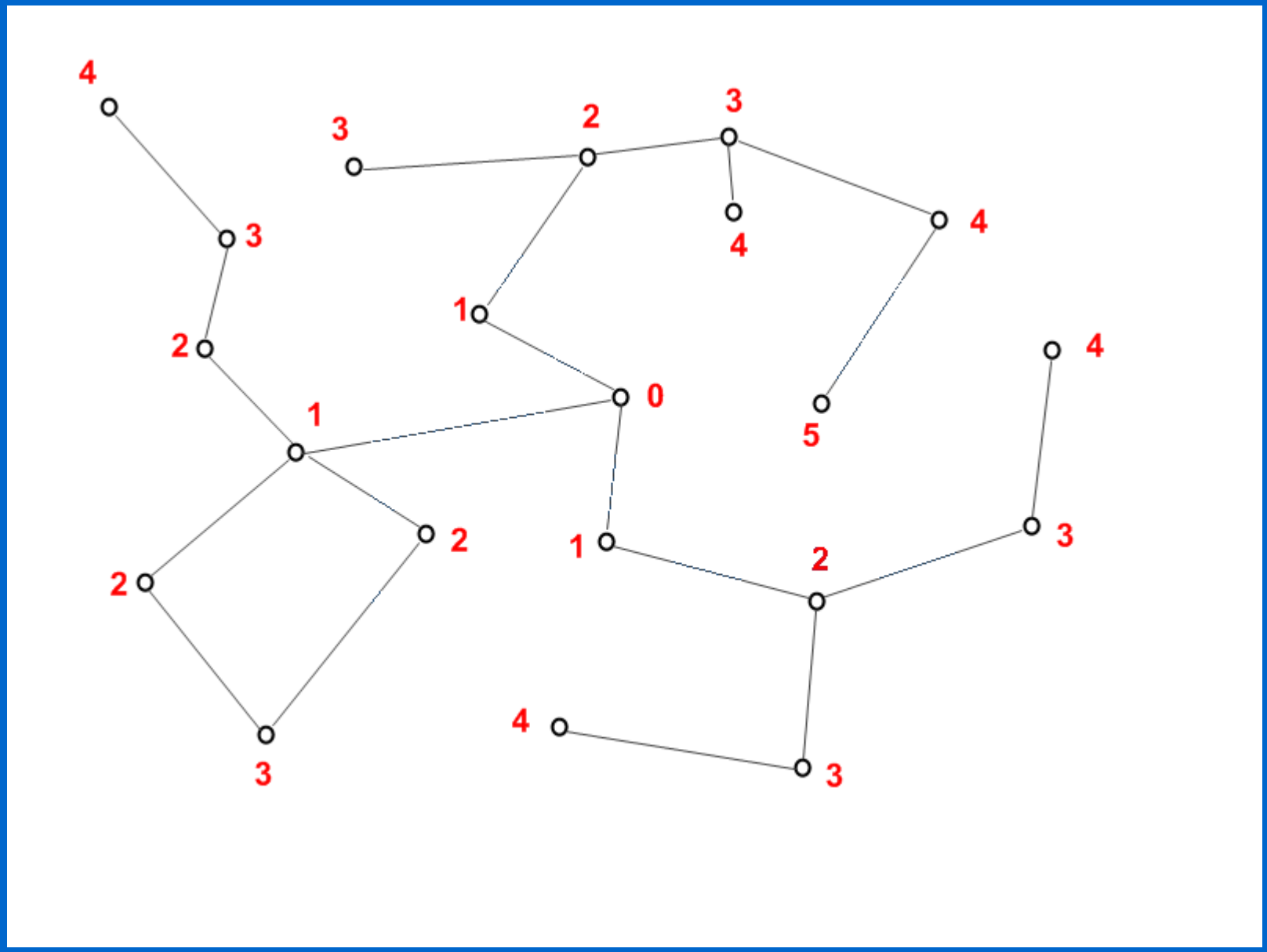
Man sieht, wie der Einsatz eines semantischen Netzes die Suche nach relevanten Dokumenten verbessern kann.

Beim eben beschriebenen Verfahren liefern nur Dokumente einen Beitrag, wo eines der Suchwörter im Dokumententext vorkommt.

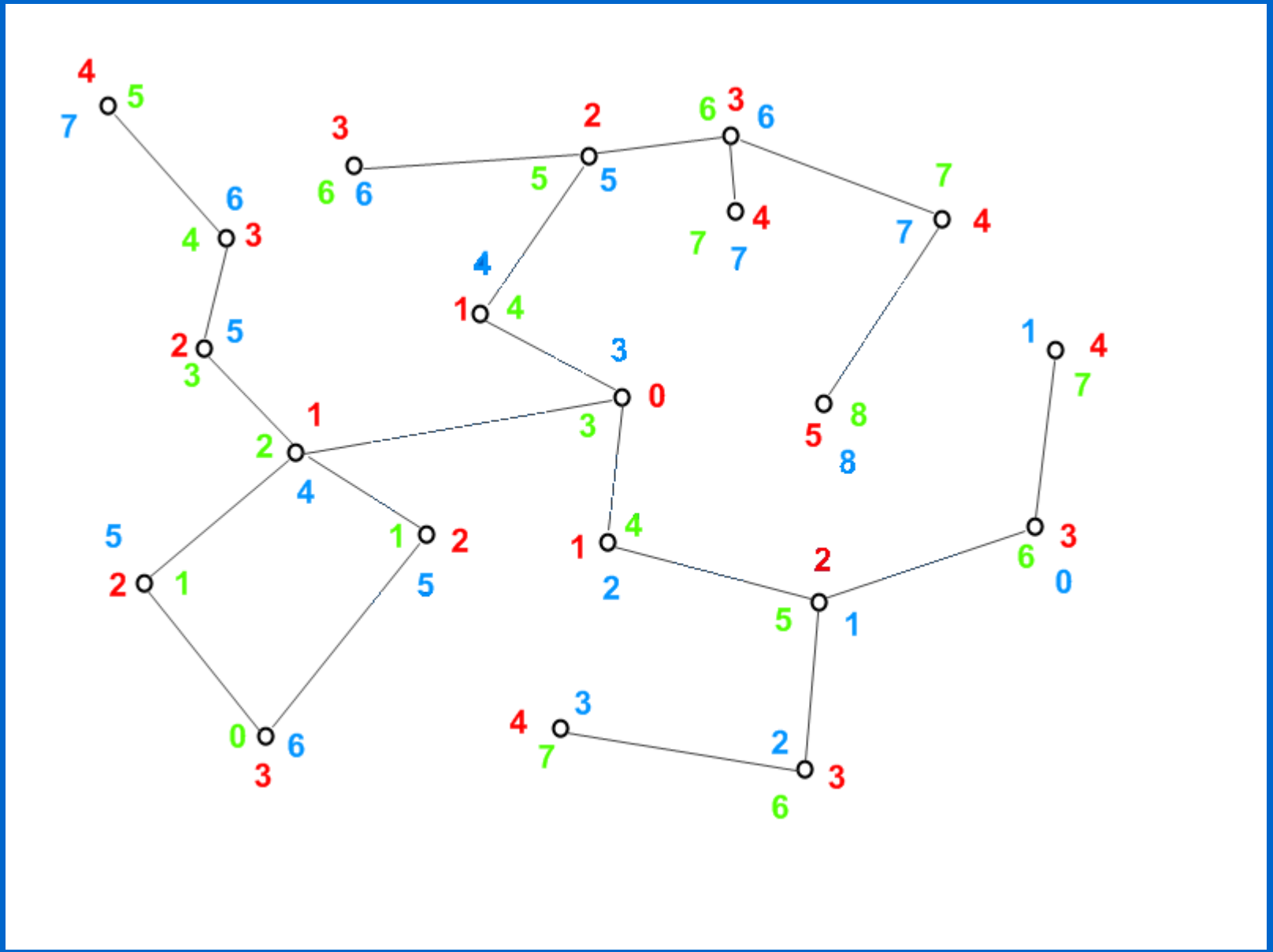
Eigentlich will man auch die alternativen Begriffe zur Suche verwenden, die das semantische Netz liefert...

... bis zu einer bestimmten maximalen semantischen Entfernung zu den ursprünglichen Suchbegriffen.

Ausgehend von einem Knoten = einem Wort werden sukzessiv Nachbarknoten numeriert, bis zu einer bestimmten maximalen Distanz



Hier im Beispiel haben wir drei Suchwörter. Das entspricht drei Startknoten im semantischen Netz. Jedes Wort stellt einen Vektor mit n Komponenten dar.



Jedes Wort stellt einen Vektor mit n Komponenten dar. Jede Komponente nennt den semantischen Abstand zu je einem Suchwort.

Um ein Dokument zu bewerten, iteriere man durch alle Wörter des Dokuments und bewahre je Komponente des Vektors den kleinsten Wert. Dies ergibt den Dokumentabstandsvektor.

Beispiel: Die Wörter eines Dokuments haben die Vektoren {(3,1,5), (2, 7, 9), (4, 8, 2)}. Dann ist der Dokumentabstandsvektor (2,1,2).

Insbesondere sind alle Komponenten 0, wenn jedes Suchwort im Dokument vorkommt, hier also (0,0,0).

Man könnte jetzt einfach die Summe der Komponenten dieses resultierenden Vektors oder die Summe der Quadrate verwenden, um Dokumente zu bewerten.

Man sieht: auch diese Technik, die mit semantischen Netzen arbeitet, führt zu einer Verbesserung der Suche.

Ein semantisches Netz liefert zwei Werkzeuge:

- Bestimmung der inneren Struktur einer Suchanfrage. Für je zwei Suchwörter wird ermittelt, ob sie sich im gleichen Cluster befinden oder nicht. Diese Bedingung ist stets mehr oder weniger erfüllt, weswegen es sich um unscharfes Clustering handelt.

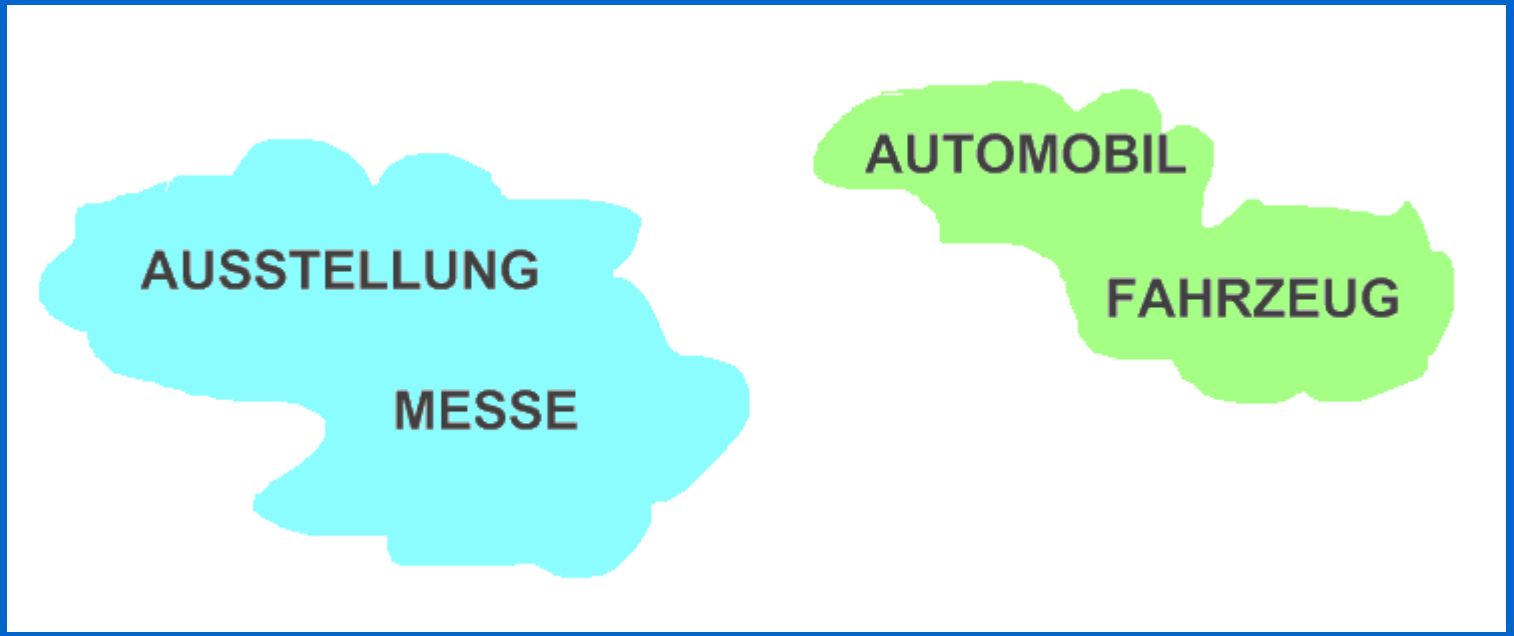
  (Den Abstand kann man auf das Intervall (0,1)
  normieren, bezogen auf den größten gefundenen semantischen Abstand.)

- Bereitstellung alternativer Suchwörter. Jedes hat zu den ursprünglichen Suchwörtern eine gewisse semantische Distanz.

Im Beispiel gibt es zwei semantische Cluster. Es gilt folgende Heuristik:

- Sind zwei Suchbegriffe aus dem gleichen semantischen Cluster, so ist eine ODER-Verknüpfung erwünscht. Es interessiert die kleinere Komponente des Dokumentabstandsvektors.

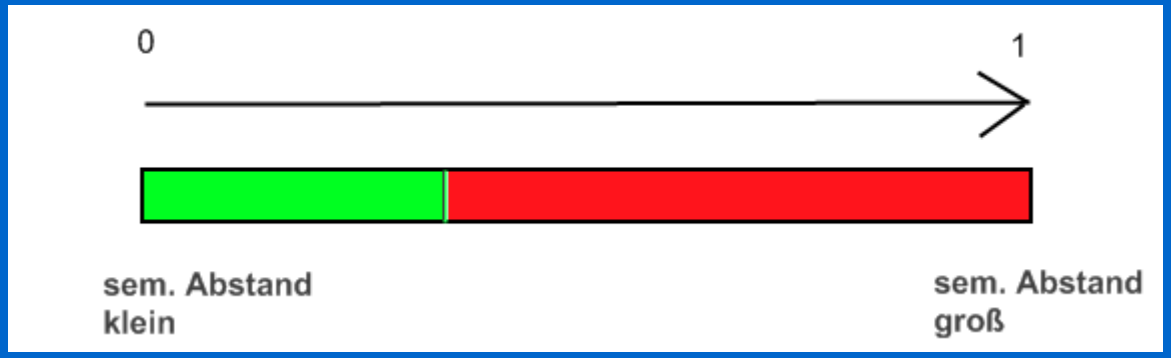
- Sind zwei Suchbegriffe aus unterschiedlichen semantischen Clustern, so ist eine UND-Verknüpfung erwünscht. Es interessiert die größere Komponente des Dokumentabstandsvektors.



(AUSSTELLUNG or MESSE) and (AUTOMOBIL or FAHRZEUG)

Der semantische Abstand ist, insbesondere bei mehr als zwei Suchbegriffen, nicht einfach klein oder groß, sondern kann eine Zwischenposition einnehmen, bezogen auf die größte semantische Distanz, die wir = 1 setzen (Normierung).

Hier sei die semantische Distanz zweier Suchbegriffe 1/3.



Für diese Suchbegriffe sollte man zu 2/3 die minimale Komponente des Dokumentabstandsvektors nehmen
und zu 1/3 die maximale.

Bewertung der Relevanz von Dokumenten

Um ein Dokument zu bewerten, ermittle den Dokumentabstandsvektor. Für alle Paare S_i, S_j von Suchbegriffen tue folgendes:

- Nimm die Komponenten x_i, x_j des Dokumentabstandsvektors.
- Ist $D(S_i, S_j)$ klein, so sind S_i und S_j ähnlich, d.h. Der Anteil ist $\min(x_i, x_j)$.
- Ist $D(S_i, S_j)$ groß, so sind S_i und S_j verschieden, d.h. Der Anteil ist $\max(x_i, x_j)$.

Der Abstand $D(S_i, S_j)$ ist nicht nur entweder klein oder groß, sondern auch irgendetwas dazwischen. Es handelt sich um einen unscharfen (fuzzy) Wahrheitswert. Entsprechend liegt der Beitrag eines Wortpaares an der Bewertung entweder näher am minimalen oder am maximalen Wert der zwei Komponenten des Dokumentabstandvektors.

Die Bewertung des Dokuments ergibt sich zu

$$\sum_{i<j} (1-D(S_i, S_j)) * \min(x_i, x_j) + D(S_i, S_j) * \max(x_i, x_j)$$

Zusammenfassung

- Semantische Netze können die Suche nach Dokumenten verbessern.

- Die Qualität der Assoziationen entscheidet über die Qualität der Suchergebnisse, d.h. eine Suche ist i.a. so gut wie das semantische Netz.

- Für jedes Fachgebiet sollte ein separates semantisches Netz erstellt werden, das die begrifflichen Zusammenhänge der jeweiligen Domäne beschreibt und die Fachausdrücke in einen assoziativen Kontext setzt.