智能问答系统实践

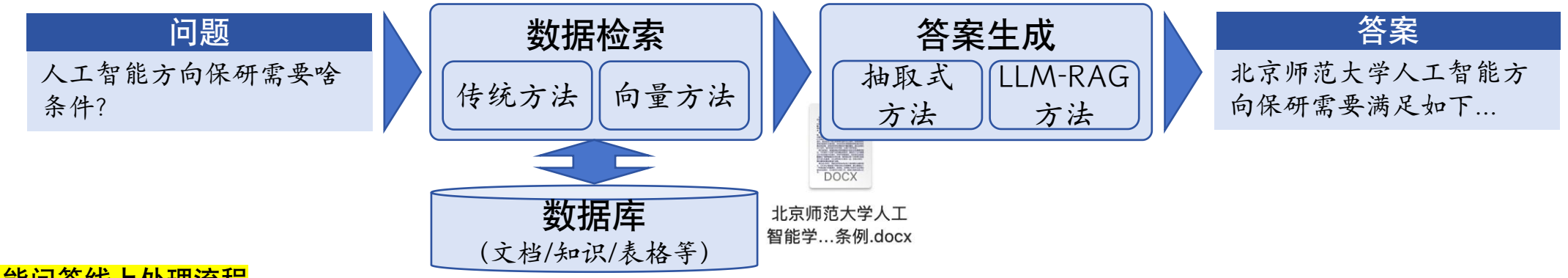
第七课：效果评估

姜文斌
北京师范大学人工智能学院
2025.04.10

我的位置

智能问答系统

针对用户提出的自然语言问题，从数据库中检索相关信息，并依据相关信息作出回答



智能问答线上处理流程

智能问答线下处理模块

问答数据库构建	数据检索模块构建	答案生成模块构建	效果评估模块构建
基于传统方法的数据建库	传统语义匹配模型构建	抽取式答案生成模型构建	文档检索效果评估
基于向量方法的数据建库	向量语义匹配模型构建	RAG式答案生成模型构建	问答整体效果评估

目录

P-R-F

- **准确率（查准率）**
  - 检索出的相关文档数占检索出文档总数的比例
  - 若检索出10篇文档，其中6篇相关，则查准率为60%

$$\text{Precision} = \frac{\text{检索出的相关文档数}}{\text{检索出的文档总数}}$$

- **召回率（查全率）**
  - 检索出的相关文档数占所有相关文档数的比例
  - 若共有20篇相关文档，检索出6篇，则查全率为30%

$$\text{Recall} = \frac{\text{检索出的相关文档数}}{\text{系统中所有相关文档数}}$$

- **F1值（F1-Score）**
  - 查准率和查全率的调和平均数，用于综合评估
  - 若查准率为60%，查全率为30%，则F1值为40%

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

MAP

- 平均准确率均值（Mean Average Precision）
  - 衡量信息检索系统排序性能的综合指标，反映多个查询中平均的检索准确性
  - 对每个查询计算平均准确率，再对所有查询取平均

- 第一步：计算单个查询的平均准确率（AP）
  - 对每个返回文档计算排序中的准确率（P@k）
  - 将相关文档P@k求和后除以相关文档总数

$$\mathrm{AP} = \frac{1}{R} \sum_{k=1}^{n} \mathrm{Precision}(k) \times \mathrm{rel}(k)$$

- 第二步：计算所有查询的MAP
  - 对所有查询的AP取平均

$$\mathrm{MAP} = \frac{1}{Q} \sum_{i=1}^{Q} \mathrm{AP}_i$$

示例

查询结果：

排名	文档ID	是否相关	Precision@k	$\mathrm{rel}(k)$	累加项
1	A	是	1.0	1	$1.0 \times 1 = 1.0$
2	B	否	0.5	0	$0.5 \times 0 = 0$
3	C	是	0.6667	1	$0.6667 \times 1 = 0.6667$
4	D	是	0.75	1	$0.75 \times 1 = 0.75$
5	E	否	0.6	0	$0.6 \times 0 = 0$

计算AP：

1. 累加相关文档的Precision：$1.0 + 0.6667 + 0.75 = 2.4167$。
2. 归一化：$R = 3$，故 $\mathrm{AP} = 2.4167/3 \approx 0.8056$。

NDCG

- 归一化折损累积增益（Normalized Discounted Cumulative Gain）
  - 衡量排序质量的指标，综合考虑文档的相关性、排序位置及用户关注度衰减
  - 评估对高相关性文档的排序能力，适用于结果列表较长且相关性分级的场景

- 第一步：计算DCG（Discounted Cumulative Gain）
  - 对每个文档，根据相关性得分和位置计算增益

$$\mathrm{DCG} = \sum_{i=1}^{n} \frac{\mathrm{rel}_i}{\log_2(i+1)}$$

- 第二步：计算归一化DCG（Normalized DCG）
  - 用理想排序的DCG（IDCG）作为基准进行归一
  - IDCG是将文档按相关性从高到低排序后计算的DCG

$$\mathrm{NDCG} = \frac{\mathrm{DCG}}{\mathrm{IDCG}}$$

示例

假设场景：

- 查询“苹果手机”，系统返回结果如下（相关性得分0–4级）：

排名	文档ID	相关性	rel_i	$\log_2(i+1)$	增益项 $\frac{\text{rel}_i}{\log_2(i+1)}$
1	A	4	4	1.0	4.0
2	B	3	3	1.58496	1.8928
3	C	2	2	2.0	1.0
4	D	0	0	2.32193	0.0
5	E	1	1	2.58496	0.3869

计算DCG：

$$\text{DCG} = 4.0 + 1.8928 + 1.0 + 0.0 + 0.3869 = 7.2797$$

理想排序（IDCG）：

- 假设相关性顺序为4, 3, 2, 1, 0，则：

排名	rel_i	$\log_2(i+1)$	增益项
1	4	1.0	4.0
2	3	1.58496	1.8928
3	2	2.0	1.0
4	1	2.32193	0.4307
5	0	2.58496	0.0

$$\text{IDCG} = 4.0 + 1.8928 + 1.0 + 0.4307 + 0.0 = 7.3235$$

计算NDCG：

$$\text{NDCG} = \frac{7.2797}{7.3235} \approx 0.994$$

指标对比

指标	适用场景	优势	局限性
查准率/查全率	二元相关性判断	直观反映检索质量	需权衡两者，无法综合评估
F1值	平衡查准率与查全率	单一指标综合评估	忽略排序信息
MAP	多查询、多相关文档场景	考虑排序与相关性	计算复杂度高
NDCG	排序敏感场景	反映用户满意度	依赖增益函数定义

目录

EM（Exact Match）

- 计算方式

$$\mathrm{EM} = \frac{\text{完全匹配的答案数}}{\text{总问题数}}$$

- 示例
  - 问题：苹果公司的CEO是谁?
  - 参考答案：蒂姆·库克
  - 预测答案1：蒂姆·库克 → EM=1
  - 预测答案2：库克 → EM=0

F1-Score

- 计算方式

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

  - Precision：正确预测词在预测答案中的比例
  - Recall：正确预测词在参考答案中的比例

- 示例
  - 参考答案：巴拉克·奥巴马
  - 预测答案：奥巴马
  - Precision=1（预测词全对），Recall=0.5（只覆盖一半参考词）
  - F1 = 2×(1×0.5)/(1+0.5) ≈ 0.67

ROUGE-L

- **核心思想**
  - 基于最长公共子序列（LCS），计算预测答案和参考答案的相似度

- **第一步：计算LCS（最长公共子序列）**
  - 基于动态规划算法，计算最长公共子序列

$$dp[i][j] = \begin{cases} dp[i-1][j-1] + 1 & \text{若 } ref[i-1] = pred[j-1] \\ \max(dp[i-1][j], dp[i][j-1]) & \text{否则} \end{cases}$$

- **第二步：计算P-R-F**

$$P = \frac{\text{LCS长度}}{\text{预测答案长度}} \qquad R = \frac{\text{LCS长度}}{\text{参考答案长度}} \qquad F1 = 2 \times \frac{P \times R}{P + R}$$

指标对比

指标	适用场景	优势	局限性
EM	抽取式问答 短答案匹配（如填空题）	严格性，直接反映答案是否完全正确 高效性，计算简单，适合大规模评估	敏感性，对微小差异敏感 无法处理部分正确情形，部分匹配时得分为0
ROUGE-L	生成式答案评估（如摘要、长文本问答）	长答案鲁棒性，适合评估长文本或摘要	计算复杂度高
F1-Score	容许部分匹配的场景（如长答案、多词回答）	平衡性，结合精确率与召回率，评估部分正确答案具有灵活性	对无关词敏感，预测答案中冗余信息可能拉低分数

目录

总结

- 评估的本质
  - 在预测答案和参考答案之间进行匹配，试图逼近人类的判断结果
  - 类比：强化学习中的奖励函数，引导系统的前进方向
- 评估的评估
  - 针对评估指标的评估，评价评估指标是否很好地拟合人类的判断
- 评估的重要性
  - 评估指标和评估数据，直接决定系统优化的方向是否是人们真正的需求
  - 警示：为什么有些系统的评估指标很高，但实际应用效果并不好？

谢谢大家！